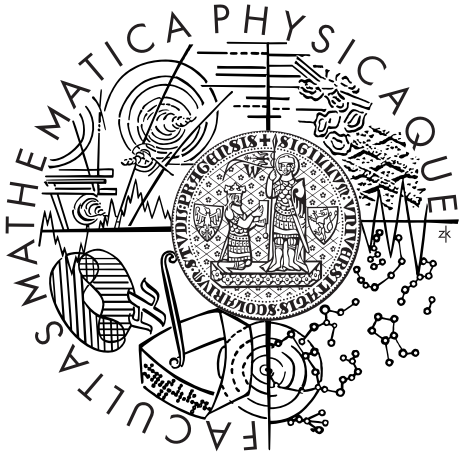
Charles University in Prague
Faculty of Mathematics and Physics

**M.Sc. Thesis**


Angelina Ivanova

# Creating a Bilingual Dictionary using Wikipedia

Free University of Bolzano
&
Charles University in Prague

Supervisor: RNDr. Daniel Zeman Ph.D.

European Masters Program in Language and Communication Technology (LCT)

Prague, 2011

# Acknowledgements

I am especially grateful to my supervisor dr. Daniel Zeman for providing me a wise supervision, helpful advices, instant replies and encouragement.

I am happy to have a chance to acknowledge the work of my coordinators from Charles University in Prague, for being in touch with me since the first year of Masters program, for being patient, kind, very attentive and helpful with all sorts of questions. I send my special thanks for help with paperwork and other questions to Ms. Libuše Brdičková and Ms. Marie Křížková.

I am very grateful to my first-year coordinator from Free University of Bolzano prof. Raffaella Bernardi and the secretary of the Computer Science faculty Ms. Federica Cumer for the readiness to answer all sort of questions and for making the adaptation to the European educational system quick and smooth.

I would like to thank all the coordinators and organizers of the LCT-Master for doing a great job and running this educational program. I would like to acknowledge the professionalism of prof. Valia Cordoni and Ms. Barbara Pernice and to thank them very much for all the help and support during these two years.

It was also great to keep in touch with the computational linguists from Russia prof. Valery Rubashkin and Lidia Pivovarova, I am very grateful to them for all the help and care.

I would also like to thank a lot dr. Gosse Bouma for the help in my Master Thesis topic definition and for being open for advice.

I also feel very lucky to meet the LCT colleagues with whom we share the deep interest in the field of natural language processing and many special memories from the time spent together. It is very exciting to live and study in the multicultural international environment.

I cannot express how important for me is the encouragement and care of my family, relatives and friends.

There are many other people to thank who influenced me very much in a very positive way such as prof. Jan Hajič, prof. Bernardo Magnini, dr. Silvie Cinková, prof. Diego Calvanese, dr. Sergio Tessaris, dr. Alessandro Artale, dr. Jan Štěpánek, dr. Zdeněk Žabokrtský, dr. Davide Martinenghi, prof. Werner Nutt, prof. Vittore Casarosa, dr. Marijke Keet, Elena Cabrio, Yashar Mehdad, Vladimír Čunát, Martin Babka and others.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No.121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

Prague, August 1, 2011

Angelina Ivanova

**Název práce:** Tvorba dvojjazyčného slovníku s použitím Wikipedie
**Autor:** Angelina Ivanova
**Katedra/Ústav:** Ústav formální a aplikované lingvistiky (32-ÚFAL)
**Vedoucí diplomové práce:** RNDr. Daniel Zeman Ph.D.
**Abstrakt:**

Strojově čitelné slovníky hrají důležitou roli v počítačové lingvistice. Osvědčily se v oblastech jako je strojový překlad a mezijazykové dobývání informací. V této práci zkoumáme kvalitu a obsah dvojjazyčných anglicko-ruských slovníků automaticky získaných z Wikipedie.

Slovníky získané z Wikipedie (zde jim pro jednoduchost říkáme Wiki-slovníky, ale tento pojem nelze zaměňovat s Wikislovníkem na `http://cs.wiktionary.org/`) se podstatně liší od tradičních slovníků: pokrytí základní terminologie z Muellerova slovníku je 7,42 %.

Pokusy s Wiki-slovníky zabudovanými do trénovacích dat pro strojový překlad vedly k malému, nicméně statisticky významnému poklesu kvality překladu ve srovnání s experimenty bez Wiki-slovníků. Předpokládáme, že hlavním důvodem je rozdílnost domén mezi slovníkem a korpusem. Na testovacích datech získaných z článků Wikipedie naopak slovník kvalitu překladu zlepšuje.

V této práci ukazujeme, jak velký je rozdíl mezi slovníky získanými ze struktury Wikipedie a tradičními slovníky. Wiki-slovníky mohou uškodit strojovému překladu, pokud se použijí pro jinou doménu. Výhoda takových slovníků je, že obsahují značné množství pojmenovaných entit a že mohou být snadno získány pro libovolný pár jazyků, které jsou zastoupené ve Wikipedii.

**Klíčová slova:** strojově čitelný slovník, Wikipedia, rozpoznávání a klasifikace pojmenovaných entit, paralelní korpus, strojový překlad

**Title:** Creating a Bilingual Dictionary using Wikipedia
**Author:** Angelina Ivanova
**Department/Institute:** Institute of Formal and Applied Linguistics (32-ÚFAL)
**Supervisor of the master thesis:** RNDr. Daniel Zeman Ph.D.
**Abstract:**

Machine-readable dictionaries play important role in the research area of computational linguistics. They gained popularity in such fields as machine translation and cross-language information extraction. In this thesis we investigate the quality and content of bilingual English-Russian dictionaries generated from Wikipedia link structure.

Wiki-dictionaries differ dramatically from the traditional dictionaries: the recall of the basic terminology on Mueller's dictionary was 7.42%.

Machine translation experiments with Wiki-dictionary incorporated into the training set resulted in the rather small, but statistically significant drop of the the quality of the translation compared to the experiment without Wiki-dictionary. We supposed that the main reason was domain difference between the dictionary and the corpus and got some evidence that on the test set collected from Wikipedia articles the model with incorporated dictionary performed better.

In this work we show how big the difference between the dictionaries developed from the Wikipedia link structure and the traditional dictionaries is. Wiki-dictionaries may reduce the quality of the machine translation if used on

the improper domain. The advantage of such dictionaries is that that they provide a great number of named entities and that they can be easily developed automatically for any language pair which is present in Wikipedia.
**Keywords:** machine-readable dictionary, Wikipedia, named entity recognition and classification, parallel corpus, machine translation

# Contents


**1 Introduction** **1**
- 1.1 Motivation . . . 1
- 1.2 Related Work . . . 1
  - 1.2.1 Tradition vs NLP Dictionaries . . . 1
  - 1.2.2 Usage in Machine Translation . . . 2
  - 1.2.3 Usage in Cross-Language Information Retrieval . . . 2
  - 1.2.4 Methods of Bilingual Dictionaries Development . . . 3
    - 1.2.4.1 Manual . . . 3
    - 1.2.4.2 Automatic from Parallel Corpora . . . 3
    - 1.2.4.3 Automatic from Comparable Corpora . . . 4
    - 1.2.4.4 Automatic from Existing Dictionaries . . . 4
    - 1.2.4.5 Automatic from Wikipedia . . . 5
- 1.3 Scope of the thesis . . . 5

**2 Dictionary** **7**
- 2.1 Wikipedia . . . 7
  - 2.1.1 Wikipedia: Structure and Markup . . . 7
  - 2.1.2 Wikipedia Resources and Tools . . . 9
    - 2.1.2.1 MediaWiki . . . 9
    - 2.1.2.2 JWPL . . . 10
    - 2.1.2.3 SQL and XML Dumps . . . 10
- 2.2 The Process of Dictionary Generation . . . 10
  - 2.2.1 Methodology . . . 10
  - 2.2.2 Filtering . . . 12
  - 2.2.3 Tokenization . . . 12
  - 2.2.4 Normalization . . . 13
- 2.3 Named Entity Recognition and Classification . . . 14
  - 2.3.1 Different Approaches to NEC Using Wikipedia . . . 15
    - 2.3.1.1 Approach That Uses First Sentence of the Article 15
    - 2.3.1.2 Approach That Uses Categories . . . 16
    - 2.3.1.3 Machine Learning Approach . . . 16
  - 2.3.2 Experiment with NERC for Wiki-dictionary . . . 17
    - 2.3.2.1 Named Entity Recognition . . . 17
    - 2.3.2.2 Named Entity Classification . . . 20
    - 2.3.2.3 Evaluation . . . 23
- 2.4 Dictionary Development for Other Language Pairs . . . 26
- 2.5 Dictionary evaluation . . . 28
  - 2.5.1 Statistics in Mueller's dictionary . . . 28
  - 2.5.2 Corpus Statistics . . . 35

**3 Evaluation on Machine Translation System** **40**
- 3.1 Machine Translation: Background . . . 40
  - 3.1.1 Statistical Approach to Machine Translation . . . 40
  - 3.1.2 Phrase-based Machine Translation . . . 41
  - 3.1.3 Machine Translation Evaluation . . . 42

3.1.3.1 Manual Evaluation . . . . . 42
3.1.3.2 Automatic evaluation . . . . . 43
3.2 Experiments with Moses . . . . . 44
3.2.1 OOV Words Analysis (UMC Test Set) . . . . . 48
3.2.2 Paired Bootstrap Re-Sampling (UMC Test Set) . . . . . 49
3.2.3 Manual Ranking (UMC Test Set) . . . . . 50
3.2.4 Data Collection for the Wiki-set . . . . . 51
3.2.5 OOV analysis (Wiki-set) . . . . . 52
3.2.6 Manual Ranking (Wiki-set) . . . . . 53
3.2.7 Evaluation Without Comments In Brackets (UMC Test Set) 55

**4 Conclusion** **56**

**Bibliography** **60**

**List of Tables** **61**

**List of Figures** **62**

**List of Abbreviations** **63**

**A Sample from English-Russian Wiki-dictionary** **64**

**B Sample from English-German Wiki-dictionary** **75**

**C Corpus statistics** **80**

# 1. Introduction

This thesis discusses the process of extraction of bilingual dictionary using Wikipedia and its impact on the performance of the statistical machine translation system.

Bilingual dictionaries are specialized dictionaries used to translate words or phrases from one language to another. There are two kinds of such dictionaries: unidirectional and bidirectional. Unidirectional dictionaries provide the translation of the meanings of the words into another language while bidirectional dictionaries allow translations to and from both languages. Such dictionaries often contain information about part of speech, gender, verb type and declension model of the words. There is a distinction between human-oriented (whether electronic or on paper) and machine-oriented dictionaries; however, dictionaries can contain both elements. Machine-oriented dictionaries usually allow to work with languages in both directions although it may be complicated by the one-to-many translations.

Wikipedia is a free, web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation[1].

## 1.1 Motivation

The bilingual dictionaries are important resources for such NLP applications as statistical machine translation and cross-language information extraction systems. They also can serve to enhance existing dictionaries, for second language teaching and learning. Manually created resources are usually more accurate and do not contain noisy information in contrast to automatically learned dictionaries. Scientific community seeks for the methods to achieve similar accuracy level and broader terminology scope by automatic means.

## 1.2 Related Work

### 1.2.1 Tradition vs NLP Dictionaries

Bilingual dictionaries can have different purposes and the information in the dictionaries depends on their intended usage. Usually the dictionaries are created in order to help users understand foreign language text or to support translation process. The format of electronic bilingual dictionaries varies from user-oriented to machine-readable or provides both interfaces.

Traditional dictionaries are viewed as important resources for NLP applications. However, these dictionaries are oriented towards human users who consult the dictionary with different purposes than NLP applications. The illustrative example is given in [8]: the most important information for application usually concerns grammatical description of words, which is usually fragmentarily available in many traditional dictionaries, with such exceptions as [22]. This leads

[1] http://en.wikipedia.org/wiki/Wikipedia

us to the conclusion that usefulness of the traditional dictionaries for NLP applications is limited and there is a need of the bilingual dictionaries designed specifically for NLP tasks.

The authors of [23] pointed out the importance of the bilingual dictionaries for lexical acquisition. Automatic learning of lexical resources might help to overcome such problems as the lack of information important for NLP tasks (usually in the cases when it is intuitive for the human users) and gaps in the terminology and proper names due to the inherent productivity of language.

### 1.2.2 Usage in Machine Translation

Bilingual dictionaries are used in Machine Translation. For example, in [25] the dictionary is used in the candidate sentence pair selection process from the non-parallel comparable corpora. In order to detect relevant sentence pairs, authors pass all possible sentence pairs through a word-overlap filter which checks in the bilingual dictionary if at least half of the words in one sentence have a translation in the other sentence. In this approach good sentences could be removed due to the fact, that dictionary might not contain the necessary word pairs. On the other hand, wrong sentence pairs could be accepted in the situation, when the majority of the words in the sentence are stop-words which are almost always present in the bilingual dictionaries. Despite that content words do not match, the number of matching stop-words pairs might be enough to fulfill the threshold requirement and introduce the wrong sentence pair. Also if the dictionary is created automatically, it may contain noisy entries which affect the correctness of the sentence alignment. Authors developed a maximum entropy classifier for parallel sentence identification from candidate sentence pairs that passed filtering step. One of the general features for this classifier independent of the word alignment is the percentage of words on each side that have a translation of the other side according to the dictionary.

In [26] bilingual dictionary plays crucial role in automatic lexical transfer and rule acquisition. Such rules are exploited in the domain of rule-based machine translation with deep syntactic and semantic processing. The architecture of the machine translation system is illustrated in Figure 1.1.

Lexical transfer rules relate predicates of source language to predicates of the target language. The bilingual dictionary proposes lexical mappings in the target language for the entries in source language. Morphological analysis of the candidate translations helps to determine the morphological processes and facilitate the mappings from semantic predicates to dictionary entries. Initially the system had a BLEU score 0.04 being able to translate only 8% of inputs. When the system was enriched with automatically generated transfer rules, its BLEU average raised up to 0.13 and the coverage increased to 22.4%.

### 1.2.3 Usage in Cross-Language Information Retrieval

Cross-Language Information Retrieval (CLIR) consists of retrieving documents written in one language using queries written in another language [31]. Dictionary-based translation is commonly used in CLIR because of the simplicity and availability. However the absence of the necessary translations of words and

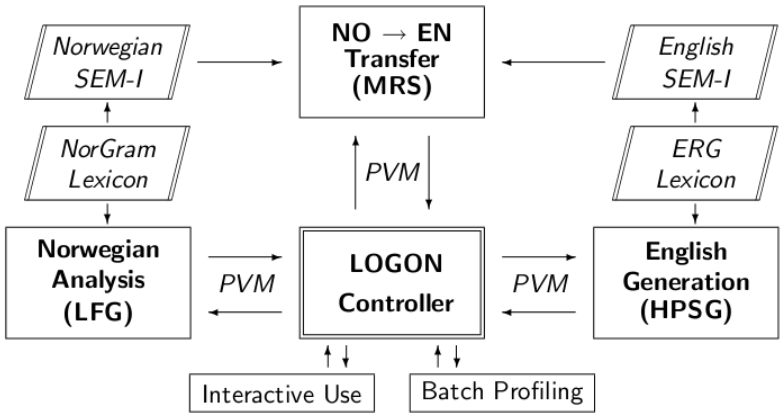



Figure 1.1: The architecture of MT system from [26]. SEM-I - Semantic Interface, MRS - Minimal Recursion Semantics, ERG - English Resource Grammar, LFG - Lexical functional grammar, HPSG - Head-driven phrase structure grammar, PVM - Parallel Virtual Machine

compounds as well as absence of the domain-specific terminology in the dictionaries cause the poor performance of CLIR systems.

### 1.2.4 Methods of Bilingual Dictionaries Development

#### 1.2.4.1 Manual

Manual dictionary making involves five principle steps: gathering of data, parsing and excerpting of entries, filing of entries according to a certain arrangement, writing of articles and publication of the final product [11]. Usually manually created resources are more accurate and comprehensive than lexicons created by automatic procedures. However, manually-built machine readable dictionaries are expensive due to the time required to create them and the effort required to cover the domain terminology and to keep the dictionaries up-to-date.

#### 1.2.4.2 Automatic from Parallel Corpora

The bilingual dictionaries can be learned automatically from parallel corpora like in [25]. The detailed description of one of the approaches to extraction of bilingual dictionary from the parallel corpora is given in [12]. In their approach the process consisted of four stages: preprocessing, alignment, extraction and filtering. The preprocessing step involved separation of the punctuation from words and lemmatization. Word alignment was done using the tool GIZA++[2]. Translation extraction was performed with the Pharaoh system tool which gave a phrase table as an output which was basically unfiltered bilingual lemma-based dictionary. The rule-based machine learning classifier Ripper was used for filtering noisy word pairs. The result dictionary was evaluated against manually built dictionary and the results showed that it was significantly more comprehensive in coverage of the terminology and had reasonable recall and high precision.

[2]http://code.google.com/p/giza-pp/

#### 1.2.4.3 Automatic from Comparable Corpora

The authors of [31] showed that bilingual dictionaries can also be learned or enriched from non-aligned comparable corpora. Unlike parallel corpora, comparable corpora are bilingual texts that are not parallel in the strict sense, but convey overlapping information[3]. Parallel corpora are available for relatively few language pairs and cover few domains, while comparable corpora is more accessible in the Web and much less expensive. The methodology proposed in [31] combined linguistic and statistical methods. Authors introduced a two-stage translation model consisting of bi-directional extraction, merging and disambiguation of the extracted bilingual terminology and application of the linguistics-based pruning. The evaluation on the large-scale text collection on Japanese-English and different weighting schemes of SMART retrieval systems proved the effectiveness of the described methods on cross-language information retrieval.

#### 1.2.4.4 Automatic from Existing Dictionaries

From a meta-lexicographic point of view the recent attempts of automatic dictionary aimed at designing reversible dictionaries in which the source language and the target language are interchangeable [14]. Reversibility becomes important in the hub-and-spoke model which contains the following steps:

1. generation of the reversible bilingual dictionary A↔B;
2. generation of the reversible bilingual dictionary A↔C;
3. establishing the links between B and C (B↔C) using pairs A↔B and A↔C.

This model was implemented in the derivation of a Danish-Finnish dictionary from the Dutch-Finnish and Dutch-Danish databases [14]. The hub-and-spoke model presumes the existence of monolingual databases with approximately the same structure and size.

The authors of [34] used pivot language to derive new bilingual dictionaries from existing ones as part of the MulTra project. They used four bilingual dictionaries that were manually created by lexicographers. Each of four dictionaries English-French, German-French, French-Italian, Spanish-French was implemented as as relation table in a relation database management system. To derive bilingual lexicon (A,C) from given (A,B) and (B,C) a relation equi-join operation was performed on the database. The validity of the obtained correspondences was checked in a parallel corpus EuroParl which means that translation pairs that were not used as translations in the corpus were filtered out. Evaluation showed that result dictionaries were of good quality but the number of corpus checked correspondences was very low due to the fact that the corpus was not large enough and probably didn't represent the general terminology well.

The authors of [8] proposed approach of automatic generation of the bilingual dictionary from a set of monolingual dictionaries and a knowledge base by enumerating the pairs of lexical entries in two monolingual dictionaries pointing to the same synonym cluster.

[3] http://research.microsoft.com/apps/video/dl.aspx?id=104110

#### 1.2.4.5 Automatic from Wikipedia

The authors of [9] propose method for creating a bilingual dictionary from interlanguage links, redirect pages and link texts. For the evaluation they extracted Japanese translations for 200 English sample terms and compared the coverage of these translations to the translations extracted from the parallel corpus. The redirect pages were used to increase the number of translation candidates. Authors introduced the notion of backward link - an incoming hyperlink of a Web page. They discovered empirically that the number of backward links of a page can be used to estimate the accuracy of a translation candidate. This can be explained by the fact that redirect pages with wrong titles or titles that are not related to the target page usually have small number of backward links. For each translation candidate they introduced the score and set up several thresholds to filter out irrelevant translation candidates. The precision and recall were better for the dictionary extracted from Wikipedia than for the dictionary extracted from parallel corpora.

The authors of [30] discuss the process of mining dictionaries from Wikipedia for under-resources languages. To achieve the goal of creation a high-quality resource they constructed parallel corpora from Wikipedia article titles, infobox information, categories, article text and dictionaries already built at each phase. The preprocessing included filtering out stop words based on term frequency across Wikipedia and conversion of the text to lower case. The dictionary building process on each step included generation of the parallel corpus and eliminating words with the help of the dictionaries from the previous steps. The final dictionary was built from dictionaries created from titles, categories, infoboxes and initial lines of the Wikipedia articles. The results were evaluated with precision, recall and F-measure and compared to the results of the other existing dictionary building systems. For the automatic evaluation it was checked if the translation of the word was the same as in any available dictionary for the language pair. For the manual evaluation native speakers marked the translation pairs as correct and wrong. Comparison with other systems showed that the method performed on the level of the existing methods of automatic generation of dictionaries from Wikipedia.

Years of research showed that machine-readable bilingual dictionaries are important resources for different NLP tasks, in particular statistical machine translation. Due to such disadvantages of the manual development of the dictionaries as the high cost and low coverage of the specific domain terminology, there is a lot of effort to investigate the possibilities of automatic dictionary creation. Wikipedia proved to be invaluable resource for bilingual dictionary extraction.

## 1.3 Scope of the thesis

We use interlanguage links and redirect pages to create the dictionary. An interlanguage link in Wikipedia is a link between two articles in different languages. The two articles linked with this type of links are usually on the same topic. A redirect page in Wikipedia contains no content but a link to another article [9]. The redirect pages usually provide synonyms, spelling alternatives or even expected typos, abbreviations or expanded text, singular or plural form of

the title of the main article.

In this thesis we seek the answers for the following questions:

- quantitative and qualitative evaluation of the dictionaries created using interlanguage links and redirect pages;
- the impact of the third language links (German) on the evaluation results for Russian-English dictionary;
- which level of accuracy we can achieve in the named entity classification task using the Wikipedia categories for Russian, English and German;
- how useful are the obtained Russian-English translation pairs within a broader NLP application such as statistical machine translation;
- if we can use the dictionary acquisition system for other language pairs.

# 2. Dictionary

## 2.1 Wikipedia

Wikipedia is a freely available multilingual, collaborative encyclopedia. The text of the articles has been written by volunteers from different parts of the world. In total it contains about 19 million articles in 282 languages. This project was launched in 2001 and its name is a combination of the word "wiki" (a technology for creating collaborative websites) and the word "encyclopedia".

### 2.1.1 Wikipedia: Structure and Markup

The body of the Wikipedia articles is formatted in Wiki markup, which is the syntax and keywords used by the MediaWiki software to format a page; the markup information tells the MediaWiki software how to display, categorize and process the article[1].

The features of the Wikipedia that are most commonly used for the research are:

1. *Entity pages* are articles that contain information focused on one single entity, such as a person, a place, or a work of art.

2. *Redirect pages* - pages which often provide equivalent names for an entity (these pages usually don't have the article body). Redirect pages are used for resolving the ambiguities. Redirections are used even for spelling resolution such as from "Microsoft's" to "Microsoft", and even for spelling correction e.g. from "Wikipeda" to "Wikipedia". These pages also solve for the abbreviation expansion such as from "PhD" to "Doctor of Philosophy". Below is the example of such page:

```
<page>
<title>Lubbock Lake</title>
<id>8029222</id>
<redirect />
<revision>
<id>88975572</id>
<timestamp>2006-11-20T08:30:36Z</timestamp>
<contributor>
<username>Hotstreets</username>
<id>588827</id>
</contributor>
<comment>[[WP:AES|←]]
Redirected page to [[Lubbock Lake Landmark]]</comment>
<text xml:space="preserve">
#REDIRECT [[Lubbock Lake Landmark]]
</text>
```

[1] http://en.wikipedia.org/wiki/Help:Wiki_markup

```
</revision>
</page>
```

3. *Disambiguation pages* - pages that link to multiple similarly named articles. Example below demonstrates one of such pages:

```
<page>
<title>Skerryvore (disambiguation)</title>
<id>8029519</id>
<revision>
<id>272999704</id>
<timestamp>2009-02-24T17:45:20Z</timestamp>
<contributor>
<username>Tawcnysc</username>
<id>2265731</id>
</contributor>
<text xml:space="preserve">
'''Skerryvore'''may have the following meanings:
*[[Skerryvore]] lighthouse in [[Argyll]], [[Scotland]].
*[[Skerryvore, Ontario]], a small community in [[Canada]] located
on the shores of [[Georgian Bay]].
* A traditional Scottish [[ceilidh]] band based in [[Glasgow]].
* Home of [[Robert Louis Stevenson]], April 1885 - August
1887, in Westbourne, [[Bournemouth]].
The building was destroyed by German bombers on the 16th
November 1940, a memorial garden is now in its place
containing a statue of the lighthouse Skerryvore.
{{disambig}}</text>
</revision>
</page>
```

4. *List pages* were built by Wikipedia contributors to group entities of the same type together. Examples of titles of such articles are "Lists of musicians", "List of sovereign states", "List of culinary fruits" etc.

5. *Infoboxes* are fixed-format tables designed to be added to the top right-hand corner of articles to consistently present a summary of some unifying aspect that the articles share and sometimes to improve navigation to other interrelated articles.

6. *Interlanguage links* - links from an article to a presumably equivalent, article in another language. In the MediaWiki format these links are marked-up the following way:

```
[[cs:Tygr džunglový]]
[[en:Tiger]]
[[fr:Tigre]]
[[it:Panthera tigris]]
```

7. *Category links* - links from an article to special "Category" pages.

```
[[Category:American psychologists]]
[[Category:American academics]]
[[Category:American educators]]
[[Category:Illinois Institute of Technology faculty]]
[[Category:University of Michigan alumni]]
```

8. *Article links* - links from one article to another of the same language. For example the sentence below contains four article links that are enclosed in double square brackets:

```
''''Present''''is the second [[studio album]] by the [[Music
of Australia|Australian]] band [[Killing Heidi]], released
in Australia on 25 October 2002 (see [[2002 in music]]).
```

The symbol "|" separates alternative name or spelling for the link and markup "'''" formats the text bold and italic. The phrase preceding the vertical bar is the name of the article, while the phrase that follows the vertical bar is the text which is actually displayed on the webpage.

### 2.1.2 Wikipedia Resources and Tools

#### 2.1.2.1 MediaWiki

The online versions of Wikipedia, Wiktionary, Wikinews and other projects run with the use of MediaWiki software. It is a free web-based wiki software application developed by the Wikimedia Foundation[2]. This software also can be used to manage a copy of Wikipedia on one's own server; it provides the user with the same convenient graphical interface for editing the encyclopedia, as online version has. This software requires a web server, PHP and a database server. We configured this software with Apache server, PHP 5 and MySQL 4.

The advantages of this tool are:

1. it lets the user browse and search his copy of Wikipedia via the same interface as the online Wikipedia;

2. the data is structured and kept in MySQL tables and the user can get access to the source data with simple SQL queries.

The disadvantage of the MediaWiki software is that the import of dumps with millions of pages into MySQL tables takes days, even on a fast server. There are different methods to import the data (from XML dumps using importDump.php script which is distributed with the MediaWiki installation package, from SQL dumps using MySQL import command etc.) but all of them turned out to be too slow for our purposes.

[2] `http://www.mediawiki.org/wiki/MediaWiki`

#### 2.1.2.2 JWPL

Another interface for Wikipedia is JWPL[3] (Java Wikipedia Library), which is a free, Java-based application programming interface that allows to access all information contained in Wikipedia. It is language-independent and provides fast and efficient access to Wikipedia. The package also includes the parser for the MediaWiki syntax. In addition to the core functionality, JWPL allows access to Wikipedia's edit history with the Wikipedia Revision Toolkit that offers tools for reconstructing past states of Wikipedia, access to all article revisions and dedicated revision storage format.

The reason why we finally decided not to work with interfaces was the efficiency issue: the scripts that work on raw data are much faster then queries through interfaces.

#### 2.1.2.3 SQL and XML Dumps

The complete copy of Wikipedia content in HTML, XML and SQL formats are available for the download on the website of Wikimedia Foundation[4].

We intended to query MySQL tables for named entity classification task. We filled in with data only two MySQL tables of Russian Wikipedia: categorylinks and page, and it took more than a week. The table page contained the titles of the articles and the table categorylinks contained the relations between the pages in the table page. The table page contained not only the titles of Wikipedia articles, but also titles of the categories, images and other files, templates, help and talk pages so we didn't need most of the data that was loaded from the SQL dump. We used only three fields of the table page: page_id, page_title and page_namespace. The field page_namespace serves to distinguish between type of pages, for example, Wikipedia articles have namespace 0 and Wikipedia categories have namespace 14. We needed just two fields of the table categorylinks: cl_from and cl_to. The field cl_from stores the page_id of the article where the link was placed. The field cl_to stores the name (excluding namespace prefix) of the category assigned to the article.

We wrote the realization of the named entity classification algorithm using MySQL tables for Russian and German Wikipedia, but the time required to load the tables and the running time of the algorithm with MySQL are too slow. We have another realization of the algorithm that works directly with XML dump and its performance is much better: it takes several hours for English Wikipedia.

## 2.2 The Process of Dictionary Generation

### 2.2.1 Methodology

The English-Japanese bilingual dictionary of [9] was developed from the link structure of Wikipedia at Osaka University. The authors described the approach and proved the advantages of this dictionary compared to manually created dictionaries. We used Interlanguage Link and Redirect Page methods proposed in

[3] `http://code.google.com/p/jwpl/`
[4] `http://dumps.wikimedia.org/`

[10] to build the dictionary. The first assumption is that the titles of the articles connected by the interlanguage link are translations of each other. The second assumption is that the titles of redirect pages are the synonyms of the title of the target page.

We created five English-Russian dictionaries:

- The baseline English-Russian dictionary was created using only interlanguage links (we will call it "I").
- The second English-Russian bilingual dictionary was created from the intersection of the interlanguage links between English, Russian and German languages (we will call it "I (EN-DE-RU)").
- The third dictionary that was used in the further experiments was created using interlanguage links and redirect pages (we will call it "I + R").
- The fourth dictionary was created from the intersection of both interlanguage links and redirect pages between English-Russian, English-German and German-Russian dictionaries of the type I+R (we will call it "I + R (EN-DE-RU)").
- The fifth dictionary was created from the the dictionary of the type I+R plus the new English-Russian links obtained by transitivity from English-German and German-Russian dictionaries (we will call it "I + R + T").

We included in the third dictionary the Russian-English translation pairs that are present in Russian Wikipedia dump and are absent from the English Wikipedia dump and the English-Russian translation pairs that are present in English Wikipedia dump and are absent from the Russian Wikipedia dump. We have such data because of two reasons: first, the dumps were made on different dates, during this gap Wikipedia editors made changes to the encyclopedia, second, some articles have only one-way mappings, e.g. there is an interlanguage link from Russian article to English article but there is no interlanguage link from this English article or any of its redirect pages to the given Russian article. For example, Russian article "Случайные знаки" has an interlanguage link to the English article "Accidental (music)". The latter article has a bi-directional interlanguage link with the article "Альтерациия (музыка)" which means it is not connected with the article "Случайные знаки" in English-Russian direction.

Some interlanguage links in Wikipedia are imprecise which results in obtaining wrong dictionary entries such as:

*24-rd Infantry Division (Germany)*
**137-я пехотная дивизия (Германия)**[5]
*Transliteration: 137-ja pexotnaja divizija (Germanija)*

An approach to overcome this issue was proposed in [24]. The problem was formalized as optimization task based on graph repair operations. The algorithm uses linear programming and a region growing techniques. The graph $G = (V, E)$ represents the union of interlanguage links provided by all editions of Wikipedia.

[5]The interlanguage link between these two articles has been deleted in the later version of Wikipedia.

### 2.2.2 Filtering

First, we replaced html entities such as & and " with their actual values, "&" and "'" correspondingly in this example. Then we filtered out the dictionary entries that contained only punctuation or only words that, lowercased, match stop words. In case only one of the synonyms matched the word from the stop-words list or a punctuation symbol, it was removed but all the rest synonyms were left in the dictionary. The motivation to delete words that match stop words was to avoid getting wrong translation results especially because we convert the data to lowercase before training language model on it. For example, the surname An in the following dictionary entry

*AN/An/An_ (surname)*
**Ан|AH**
*Transliteration: an/an*

can be wrongly interpreted as an indefinite article "an" when we do Russian-English translation.

Another example:

*On/ON/ON_ (album)/On_ (album)*
**Он**
*Transliteration: on*

We risk to get all personal pronouns "он" to be wrongly translated as "on" instead of "he".

We didn't remove any other dictionary entries though there are some candidate articles that are correct but don't seem very useful for the translation tasks, for example, such frequent types of titles as

*2003 in literature*
*2003 in rail transport*
*9 (number)*
*Czech Republic at the 2004 Summer Olympics*
*Czech Republic at the 1994 Winter Olympics*
*List of birds of Slovakia*

### 2.2.3 Tokenization

We created the version of the Wiki-dictionary with tokenized phrases for further corpus evaluation. For the tokenization task we used tokenizer which is provided in Europarl v6 Preprocessing Tools suit written by Philipp Koehn and Josh Schroeder[6]. It is written in Perl and uses regular expressions of this programming language. This tokenizer requires the dictionaries of the non-breaking prefixes. The authors define such prefix as any word ending with a period that does not indicate the end of the sentence. The typical example for English is Mr. and Mrs. Such example list for English language is provided with the source code of the tool. The non-breaking prefixes list for Russian language is available

[6] `https://victorio.uit.no/langtech/trunk/tools/alignment-tools/europarl/`

on the Corpus-Tools project web page[7]. Tokenizer does not separate the prefixes listed in the non-breaking prefixes dictionary from the period with a space. There is also a special case of prefixes called NUMERIC_ONLY. These prefixes are treated as non-breaking only before numbers, e.g. in the sentence "Article No. 24" the No. functions as non-breaking prefix, while in the sentence "No. I disagree with you" No should be treated as a word.

### 2.2.4 Normalization

Russian language has rich inflection system therefore we in further corpus evaluation experiments we were interested to obtain not only token statistics of occurrence of the dictionary translation pairs in the corpus, but also statistics of of occurrence the normalized dictionary translation pairs in the corpus. Therefore we created the version of the Wiki-dictionary with normalized translation pairs. The normalization tool was run on the tokenized version of the Wiki-dictionary.

One of the suitable tools for this task could be TreeTagger[8] which is a free software for annotating text with part-of-speech and lemma information.

Dialing Morphology[9] is another publicly available morphological analyzer which works with inputs in Russian, English and German. This tool can output word's normal form, morphological features and the full paradigm. The tool can successfully predict all the features of the words that are not in in its dictionary.

We will use word "sacrilicious" (blend of sacrilegious and delicious) in the example:

```
1. Lemma: SACRILICIOUS
Morphological features: noun, common noun, singular
Paradigm: sacrilicious (singular), sacriliciouses (plural)
2. Lemma: SACRILICIOUS
Morphological features: adjective
Paradigm: sacrilicious
```

However, this tool has some difficulties with the named entities in English language. For example, it can only recognize that the surname "Djokovic" is a proper noun but it cannot disambiguate if it is singular or plural.

We chose a tool which is based on the Dialing Morphology tool and is handy for a task of normalization - Lemmatizer[10]. The software works with English and Russian in UTF8 encoding without conversion to Windows-1251 for Russian or to Latin-1 for English. The output includes paradigms and morphological information (part of speech, case, gender, tense etc) for the input word. It is dictionary-based, however, it can also predict paradigms and morphological features for the out of vocabulary words. Software is written in C and C++ and works on Unix-style operational systems; it can be accessed via C++ wrapper. We wrote a script that runs lemmatizer on our data and leaves the words for which it cannot predict the normal form in their original form.

[7] http://code.google.com/corpus-tools/source/browse/trunk/Lingua-Sentence

[8] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[9] http://www.aot.ru/demo/morph.html

[10] http://www.lemmatizer.org

For example, for the tokenized sentence from the training set

```
If that happens , it is likely that the Serbs of North Kosovo would
declare their own independence .
```

the lemmatizer produced the following output:

```
IF THAT HAPPEN , IT BE LIKELY THAT THE SERB OF NORTH KOSOVO WOULD
DECLARE THEIR OWN INDEPENDENCE .
```

The corresponding tokenized Russian sentence from the training set

```
Если это произойдет , то существует вероятность того , что сербы
в Северном Косово провозгласят собственную независимость .
```
*Transliteration: esli èto proizojdet , to suščstvuet verojatnost togo , čto serby v severnom kosovo provozglasjat sobstvennuju nezavisimost .*

is processed with the following output:

```
ЕСЛИ ЭТО ПРОИЗОЙТИ , ТО СУЩЕСТВОВАТЬ ВЕРОЯТНОСТЬ ТОГО , ЧТО СЕРБ
В СЕВЕРНЫЙ КОСОВО ПРОВОЗГЛАСИТЬ СОБСТВЕННЫЙ НЕЗАВИСИМОСТЬ .
```
*Transliteration: esli èto proizojti , to suščstvovat verojatnost togo , čto serb v severnyj kosovo provozglasit sobstvennyj nezavisimost .*

## 2.3 Named Entity Recognition and Classification

In this study we are interested in examining the content of the dictionary that we built in the previous section. The main question that we would like to answer is how many of the dictionary entries are names and how these names are distributed among the standard classes of named entities. This motivates us to work on the problems of named entity recognition and classification.

Named entity recognition (NER) is the task of processing of the text with the goal to identify the names among other words.

Named entity classification is the task of assigning the name to the particular class.

These two modules of text processing are nowadays present in various NLP applications, such as systems for Information Extraction, Information Retrieval, Summarisation and Question Answering. The term "named entity" was coined for the Sixth Message Understanding Conference (MUC-6). In the last few years Wikipedia has been used as an important resource of the named entities and there were a number of experiments of transforming Wikipedia to a named entity-annotated corpus.

The systems for named entity recognition and classification (NERC) are usually evaluated based on how their output compares with the output of human linguists. In the CONLL 2003 shared task on language-independent named entity recognition described in [33] a simple evaluation protocol was applied: the NERC systems were compared by F1-score with the precision being the percentage of named entities found by the system that are correct and the recall being

the percentage of named entities present in the solution that are found by the system. A named entity is correct only if it is an exact match with the corresponding entity in the solution.

In this study we use the four standard measures that are defined for the two-class confusion matrix (see Table 2.1): accuracy, precision, recall and F1-score (see Table 2.2).

| | | **Predicted label** | |
|---|---|---|---|
| | | **True** | **False** |
| **Actual label** | **True** | True Positive (TP) | False Negative (FN) |
| | **False** | False Positive (FP) | True Negative (TN) |

Table 2.1: Confusion matrix

| **Metric** | **Definition** | **Description** |
|---|---|---|
| Accuracy | A=(TP+TN)/(TP+TN+FP+FN) | percentage of correct classifications |
| Precision | P=TP/(TP+FP) | percentage of correct positive classifications |
| Recall | R=TP/(TP+FN) | percentage of positive cases correctly classified as positive |
| F1-score | F=2PR/(P+R) | harmonic mean of precision and recall |

Table 2.2: Accuracy, Precision, Recall, F1-score

### 2.3.1 Different Approaches to NEC Using Wikipedia

Our second step in examining the dictionary content is the classification of the named entities. First, we look into the existing approaches to this problem.

#### 2.3.1.1 Approach That Uses First Sentence of the Article

For example, the authors of [15] retrieve the Wikipedia article for each candidate word sequence and extract a category label from the first sentence of the article. They work under the assumption that the first sentence of the article is the definition of the title word sequence. The authors used a noun phrase following forms of the verb "to be" to derive a label. For example, for the name "Freddy Mercury" their method would first find the corresponding Wikipedia article, then it would extract the first sentence: "Freddie Mercury (born Farrokh Bulsara, 5 September 1946 – 24 November 1991) was a British musician, best known as the lead vocalist and a songwriter of the rock band Queen" and from this sentence it would extract the category label "musician". The authors used these labels as features in a CRF-based NE tagger.

#### 2.3.1.2 Approach That Uses Categories

The authors of [29] extracted the categories to which the entry was assigned for each article title (excluding Template pages, Wikipedia administrative pages, and articles whose title begins with "List of"). They derived a set of key phrases for each named entity type manually. The most important key phrases are shown on the Figure 2.1

| | |
|---|---|
| PERSON | "People by", "People in", "People from", "Living people", "births", "deaths", "by occupation", "Surname", "Given names", "Biography stub", "human names" |
| ORG | "Companies", "Teams", "Organizations", "Businesses", "Media by", "Political parties", "Clubs", "Advocacy groups", "Unions", "Corporations", "Newspapers", "Agencies", "Colleges", "Universities" , "Legislatures", "Company stub", "Team stub", "University stub", "Club stub" |
| GPE | "Cities", "Countries", "Territories", "Counties", "Villages", "Municipalities", "States" (not part of "United States"), "Republics", "Regions", "Settlements" |
| DATE | "Days", "Months", "Years", "Centuries" |
| NONE | "Lists", "List of", "Wars", "Incidents" |

Figure 2.1: Sample key category phrases from [29]

For each article, the algorithm searched the hierarchy of categories until a threshold of reliability was passed or a preset limit of how far to search was reached. The entity was classified in the case when at least one of the categories in the searched hierarchy matched contained a key phrase. Finally, the authors used Wiktionary to eliminate some common nouns. For example, "Tributary" is an entry in Wikipedia which would be classified as a Location from the category structure. However, it is found as a common noun in Wiktionary.

To categorize the terms in languages other than English they applied two techniques to make a decision based on English language information. If the title of the Wikipedia page that describes the term is connected to the corresponding article in the English language via interlanguage link, the title in the foreign language is assigned the same type as the English title. If the article doesn't have English equivalent, they attempt to make a decision based on the category information searching for the English equivalents of the categories in the foreign language. This proofed to be effective because many of the most useful categories have equivalents in different languages.

#### 2.3.1.3 Machine Learning Approach

The paper [7] reports about a study on the classification of Wikipedia pages with named entity labels. The 5-fold classification (PER - person; MISC - mis-

cellaneous; ORG - organization; LOC - location; COMM - common object, not a named entity) was performed using two machine learning algorithms: CVN and Naive Bayes. They collected page-based and context features. Page-based feature vectors were collected from the XML syntax of Wikipedia articles. Context feature vectors use a text window of three words to the left and to the right of the reference.

The group of page-based features includes:

**Bag of Words (BOW)** - the term frequency representation of the entire page.

**Structured Data (STRUCT)** - the bag-of-words feature vector that corresponds to the structured data only (such as data from the tables).

**First Paragraph (FPAR)** - bag-of-word representation of the first paragraph of the article.

**Abstract (ABS)** - bag-of-word feature vector based on the abstract of the article.

**Surface Forms and Disambiguations (SFD)** - feature vectors that represent the surface forms of the article links that are present in the text of the article and their disambiguated meanings.

The group of context features includes:

**Unigram Context (UCON)** - the vector, in which each feature represents the total number of times that a term $t$ appears in position $i$ around the entity.

**Bigram Context (BCON)** - the bigram-based model similar to UCON.

| | Precision | | Recall | | F-measure | |
|---|---|---|---|---|---|---|
| | SVM | NB | SVM | NB | SVM | NB |
| BOW | **0.901** | 0.858 | 0.894 | 0.880 | **0.897** | 0.869 |
| SFD | 0.851 | 0.775 | 0.830 | 0.882 | 0.840 | 0.825 |
| STRUCT | **0.888** | 0.840 | 0.875 | 0.856 | **0.881** | 0.848 |
| FPAR | 0.867 | 0.872 | 0.854 | 0.896 | 0.860 | **0.884** |
| ABS | 0.861 | 0.833 | 0.852 | 0.885 | 0.857 | 0.858 |
| BCON | 0.311 | 0.245 | 0.291 | 0.334 | 0.300 | 0.283 |

Figure 2.2: Average precision, recall and F1 measure values for the multi-class NERC task from [7]. Results are obtained using SVM and Naive Bayes.

The summary of the most important results is shown on the Figure 2.2. The best results were for BOW features. The authors emphasized the fact that STRUCT features proofed to be distinctive for the task of identification of the type of the page while the bigram contextual features BCON gave surprisingly low results, even in a co-training setup.

### 2.3.2 Experiment with NERC for Wiki-dictionary

#### 2.3.2.1 Named Entity Recognition

For the NER task we used the heuristics proposed in [5]:

1. If the title of the article is a multi-word expression, the algorithm checks the capitalization of all content words, e.g. words other than prepositions, determiners, conjunctions, relative pronouns or negations. The title is recognized as a named entity only if all the content words are capitalized. The list of the non-content words used in our experiments is shown on the Figure 2.3. We collected stop-words from several sources: the site dedicated to the Snowball language[11], the website Ranks Webmaster Tools[12], the blog Armand Brahaj[13] and the website Text Fixer[14].

2. If the title is one word that contains at least two capital letters, than it is recognized as a named entity.

3. Algorithm counts how many times the title occurs in the positions other than the beginning of the sentence in the text of the article. If at least 75% of times this title is capitalized, then it is recognized as a named entity.

| | | |
|---|---|---|
| 'twixt, a, abaft, aboard, about, above, absent, across, afore, after, against, all, along, alongside, although, amid, amidst, among, amongst, an, and, anent, anny, anoda, another, anti, any, any-and-all, apropos, around, as, aside, astride, at, athwart, atop, barring, beaucoup, because, before, behind, behither, below, beneath, beside, besides, between, betwixt, beyond, both, but, by, c., ca., certain, chez, circa, concerning, considering, cum, dat, despite, dis, down, during, each, either, | enough, enuf, enuff, eny, ere, euerie, even, everie, every, except, excluding, failing, few, fewer, fewest, fewscore, following, for, fornenst, fornent, from, given, hevery, how, however, if, in, including, inside, into, last, least, like, little, many, mid, midst, minus, more, mos', most, much, near, neither, next, nil, no, none, nor, not, notwithstanding, now, of, off, on, once, only, onto, opposite, or, other, out, outside, outwith, over, overmuch, own, pace, past, per, plus, pro, provided, qua, re, | regarding, round, sans, save, severall, since, so, some, such, than, that, the, them, these, thine, this, those, though, through, throughout, thru, thruout, till, times, to, toward, towards, umpteen, under, underneath, unless, unlike, until, unto, up, upon, v., versus, via, vice, vis-à-vis, vs., w/, w/o, what, whatever, when, whenever, where, whereas, wherever, whether, which, whichever, while, who, whoever, whom, whomever, whose, with, within, without, worth, yet, yonder |

Figure 2.3: Words that were considered non-content in the NER experiment

We introduced two more features that allow us to consider a title of the Wikipedia article as a named entity:

- If the title is one word and contains at least one capital letter and at least one digit, then it is a named entity.

- If the title is a multi-word expression, each word of which starts with the capital letter *o*r digit, then it is a named entity.

The first feature allows us not to miss the names of the comets, meteorits etc. such as "1P/Halley", "9P/Tempel". The second feature lets us to find such named entities as "1955-56 European Cup" and stand-alone dates such as "16 BC".

[11] `http://snowball.tartarus.org/algorithms/russian/stop.txt`
[12] `http://www.ranks.nl/resources/stopwords.html`
[13] `http://armandbrahaj.blog.al/2009/04/14/list-of-english-stop-words/`
[14] `http://www.textfixer.com/resources/common-english-words.txt`

We applied the heuristics above to the titles of English Wikipedia and translated results for the titles of the Russian Wikipedia. The results of NER on the I+R Wiki-dictionary (see section 2.2.1 for the dictionary types definitions) are summarized in the Table 2.3, that show that 88% of the Wiki-dictionary entries were considered named entities and 12% of the Wiki-dictionary entries were considered common words.

| | |
|---|---|
| Total number of entries in the dictionary | 348405 |
| Total number of entries recognized as NE | 305514 |
| Total number of entries recognized as non-NE | 42891 |

Table 2.3: NER of the Wiki-dictionary entries

We randomly collected 100 entries from the dictionary and manually evaluated if the named entity recognition was done correctly for each entry. The results from Table 2.4 show that we achieved only 74% accuracy.

| | |
|---|---|
| Size of the Wiki-dict | 348405 |
| Size of the Sample | 100 |
| # of correctly classified entries | 74 |
| # of wrongly classified entries | 26 |
| True Positives | 69 |
| True Negatives | 5 |
| False Positives | 20 |
| False Negatives | 6 |

Table 2.4: Evaluation of NER on the sample from Wiki-dictionary

This sample reveals the problems of the heuristics: it tends to assign the named entity flag to the terms and phrases that are not named entities. The main reason for this is that heuristic that says to treat the multi-word title as a named entity in case all the words are capitalized, is not always true. There are quite many cases when all the words of the multi-word special term are capitalized:

*Museum of Fine Arts | Museum of fine arts | Musée des Beaux-Arts | Musee des Beaux Arts | Musée des Beaux Arts | Musee des Beaux-Arts | Musée de Beaux Art*
**Музеи изобразительных искусств | Музей изобразительных искусств | Музеи изящных искусств | Музей изящных искусств**
*Transliteration: Muzei izobrazitelnyx iskusstv | Muzej izobrazitelnix iskusstv | Muzei izjaščnyx iskusstv | Muzej izjaščnyx iskysstv*

*Cloning | Klone | Cloning Extinct Species | Cloned | Clone cells | Cell cloning | Cloner*
**Клонирование (биотехнология) | Клон (биология) | Клон (биотехнология)**
*Transliteration: Klonirovanie (biotexnologija) | Klon (biologija) | Klon (biotexnologija)*

*Cuban Solenodon | Almiqui | Solenodon cubanus | Atopogale cubana | Atopogale*
**Кубинский щелезуб | Solenodon cubanus | Solenodon cubana**
*Transliteration: kubinskij ščelezub*

"Museum of Fine Arts" is just a general term for the particular type of the museum, "cloning" is a biomedical term that names the biotechnology and "cuban solenodon" is a biological term that classifies particular specie; non of these terms is a named entity.

The second type of false positive are the phrases that contain named entities such as

*Pakistan at the 1948 Summer Olympics*
**Пакистан на летних Олимпийских играх 1948**
*Transliteration: pakistan na letnix olimpijskix igrax 1948*

*Saturn Award for Best Make-up*
**Премия «Сатурн» за лучший грим**
*Transliteration: premija "saturn" za lučšij grim*

The heuristic that says that a word is a named entity if it contains two or more capital letters, is not always true:

*Web conferencing | Webinar | WebConferencing | Web conferencing appliance*
**Онлайн-семинар | Вебинар (онлайн-семинар) | Вебинар**
*Transliteration: onlajn-seminar | vebinar (onlajn-seminar) | vebinar*

False negatives occur because some articles describing one-word named entity are very short and the name doesn't occur on the positions other than beginning of the sentence in the article, for example:

*Muaná*
**Муана**
*Transliteration: muana*

The second reason for getting false negatives is that some of the pairs occurred only in the Russian Wikipedia so for such pairs we only ran the heuristics for English that don't utilize the text of the article. We have such pairs because the dumps were created on different dates and some interlanguage links present in the Russian Wikipedia dump are not in the English Wikipedia dump and vice versa.

#### 2.3.2.2 Named Entity Classification

We used the named entity classes from Conference on Computational Natural Language Learning (CoNLL) 2003 Shared Task shown in Table 2.5

| | |
|---|---|
| PER | person |
| LOC | location |
| ORG | organization |
| MISC | miscellaneous |

Table 2.5: Named entity types

**Animated Entities (PER)** type is for animate entities can be either of type human or non-human. Human entities are either humans that are known to have lived (e.g., "Anna German", "Milan Kundera", "Antonín Dvořák") or humanoid individuals in fictional works, such as books, movies, TV shows, and comics (e.g., "Mowgli", "Snow White", "Cinderella") and mythological figures (e.g. "Zeus", "Heracles", "Cupid"). Non-human entities are any particular animal or alien that has lived or that is described in a fictional work and can be singled out using a name (e.g. "ALF").

**Organization Entities (ORG)** type is for the organization entities that have some formally established association (e.g. "Czech Airlines", "Czech Football Association", "Komerční banka").

**Location Entities (LOC)** type is for the physical locations (regions in space) defined by geographical, astronomical, or political criteria ("Czech Republic", "Russia", "Mount Kilimanjaro").

**Miscellaneous Entities (MISC)** - events, works of art, artifacts ("Olympic Games", "Eat, Pray, Love", "The Lord of the Rings").

We could classify only the titles of the entity pages because redirect pages are not included in the categories. Due to these reasons, from each entry of the dictionary that was recognized as named entity we took one English named entity and one Russian named entity (those, that are obtained from the titles of the entity pages) for the process of classification. Therefore we had 305,514 named entities for English language to classify and 305,514 named entities for Russian language to classify.

We implemented bootstrapping algorithm described in [17] for the named entity classification task.

Firstly, we manually chose a set of seed categories for each type and classified all pages placed in the seed categories of corresponding type. Articles that were classified as more than one type were filtered out because we introduced the restriction that only one type can be assigned to each named entity. We collected all the categories of the classified named entities and put them into the vectors of corresponding type.

The second step was bootstrapping. We compared the vector of categories of each unclassified named entity with the four vectors of categories that corresponded to four named entity types. We used the dice similarity measure to assign the most appropriate type to a given named entity. The weight of the category in the vector of categories was the number of classified named entities of this type that were placed in this category. When we calculated the dice similarity between the vector of categories of unclassified named entity and the vector of categories of one of the types, we assigned weight 1 to all the categories that occurred in the type vector and weight 0 to all the categories that didn't occur in the type. The algorithm ran in 10 iterations, performing classification of 10% of the whole list

of named entities at each iteration.

The Table 2.6 shows that the algorithm classified more Russian named entities than English named entities. As the algorithm depends on the choice of the seed categories, we suppose that this is the reasons why the number of classified Russian named entities prevail.

| # of entries recognized as NE | 305,514 |
|---|---|
| # of classified English named entities | 259,417 |
| # of classified Russian named entities | 274,146 |

Table 2.6: NEC of the named entities from the Wiki-dictionary

After we performed named entity classification for two languages, we had a choice:

- Intersect the labels for English and Russian. If the labels matched, the whole entry would be classified with this label. If the labels didn't match, the entry gets the label "unclassified".
- Translate the labels from the results of the named entity classification for English, ignoring the classification for Russian.
- Translate the labels from the results of the named entity classification for Russian, ignoring the classification for English.

We created all three versions of the dictionary and evaluated each of them.

First, we noticed the such problems as the names of films were classified as organizations in all three versions of the dictionary. In order to improve the results of the named entity recognition and classification, we collected the comments in brackets from the titles, sorted them by frequency and examined 100 most frequent comments for English and 100 most frequent comments for Russian. Most of them are useful for the tasks of NERC, e. g. "singer", "company", "city", "album".

However, some of these comments cannot be helpful for the distinguishing the classes of the named entities, for example, "music", "mathematics", "1941".

Also, some comments such as "people", "unit" indicate that a term is not a named entity.

We collected some additional comments during the manual evaluation. The Figure2.4 shows the comments for English language grouped by classes.

We noticed that the comments like "2006 film" are very common that is why for each title in the dictionary entry we additionally checked if it had the comment that contained the word "film" or the word "movie". We also checked the titles themselves and removed named entity label in case the title was like "Czech Republic at the 1994 Winter Olympics" or "List of colors", otherwise, we additionally checked if the title contained words "olympic games", "cup", "league" or "championship" which are the key words for MISC class. All these comparisons were done in lower case.

We applied the heuristics based on the comments in brackets not to individual named entities but to the whole entries because the task of the named entity classification was performed only for the titles from the entity pages (the redirect

| Category | Comments |
|---|---|
| **PER** | footballer, mythology, musician, surname, character, singer, name, actor, politician, artist, racing driver, racecar driver, clarinetist, conductor, theater director, violinist, pianist, chess player |
| **ORG** | company, agency |
| **LOC** | district, city, river, town, province, aargau, berne, bern, municipality, graubünden, graubunden, region, graubuenden, zurich, state, zürich, zuerich, island, paris metro, moscow metro, basel-land, basel-country, thurgau, grisons, solothurn, lucerne |
| **MISC** | film, album, song, band, movie, number, video game, novel, tv series, software, series, comics, book, computer game, programming language, magazine, episode, ship, chess opening, manga, tv series, single, 2006 film, 2008 film, 2009 film, year, tv miniseries, lost, moon, constellation, the outer limits, trigraph, game |
| **Not a named entity** | people, unit, measurement, language, genus |

Figure 2.4: Comments extracted from the titles of Wikipedia articles used to correct the results of NERC for Wiki-dictionary

pages do not have categories), as we already mentioned before. However, the comments in brackets occur in the titles of the interlanguage links as well as in the titles of the redirect pages. The Table 2.7 shows the changes in the classification of the named entities after applying the heuristics to the three versions of the dictionary.

The number of entries that had "unclassified" label and are now classified is the highest for the version of the dictionary that was created from the intersection of the named entity labels for English and Russian. This happened because this version of the dictionary contained the highest number of the named entities with the label "unclassified" due to the fact that labels for English and Russian often didn't match. On the contrary, the number of entries for which one class label was changed for another one is the lowest for the dictionary that was created from the intersection of the named entity labels for English and Russian. This is explained by the fact that we assigned a class label to the entry only in case named entities from both English and Russian sides had the same label. The classification that we got by intersection of the labels was more precise than in the other two versions of the dictionary.

The numbers in the table also show us that with this approach we could correct only a very small number of entries with respect to the total number of entries in the dictionary.

#### 2.3.2.3 Evaluation

To evaluate the results of the named entity classification we took a random sample of 300 entries and manually marked them up with the appropriate labels.

The true distribution of the classes of named entities on the evaluation set of the size of 300 entries was:

- 73 entities of class PER;

| | intersection of En and Ru NE labels | NE labels translated from En | NE labels translated from Ru |
|---|---|---|---|
| **Number of entries in the Wiki-dictionary** | 348,405 | | |
| **# of entries that had "unclassified" label and are now classified** | 18,005 | 6,222 | 4,589 |
| **# of entries for which one class label was changed for another one** | 9,201 | 16,788 | 15,236 |
| **# of entries that didn't have NE label but are now classified NE** | 3,267 | | |
| **# of entries that were considered NE and are now considered non-NE** | 9,139 | | |

Table 2.7: Changes in NEC of the three versions of the Wiki-dictionary after making the corrections based on comments in brackets.The first version of the Wiki-dictionary was created by intersection of the named entity labels for English and Russian; the second version was created by translation of the named entity labels of English words ignoring the labels for Russian; the third version was created by translation of the named entity labels of Russian words ignoring the labels for English

- 8 entities of class ORG;
- 88 entities of class LOC;
- 47 entities of class MISC;
- 216 named entities in total.

Then we ran the script that evaluated the three versions of the dictionary (the version with the intersection of the English and Russian labels, the version with the labels translated from the results of NEC for English and the version with the labels translated from the results of NEC for Russian) against the manually prepared set.

As we saw in the Table 2.7 the changes for the binary classification NE/non-NE were uniq for all three versions of the Wiki-dictionary after examining the comments in brackets. Therefore the general statistics about the NER task is the same for all three versions (see Table2.8).

The absolute number of the correctly classified entries (number of entries that match the entries from the manually marked-up set exactly) is for the version of the dictionary with the labels translated from Russian (see Table 2.9). As the performance of the algorithm depends on the choice of the seed categories, we think that the quality of the seed categories that we had chosen for the Russian

| | |
|---|---|
| True positives | 208 |
| True negatives | 22 |
| False positives | 62 |
| False negatives | 8 |
| Accuracy | 76.67% |
| Precision | 77.04% |
| Recall | 96.3% |
| F1-score | 85.6% |

Table 2.8: Evaluation of the NER for Wiki-dictionary on the manually marked-up set

language might be higher than the quality of the seed categories chosen for the English language.

| | **intersection of En and Ru NE labels** | **NE labels translated from En** | **NE labels translated from Ru** |
|---|---|---|---|
| **# of entries in the sample** | 300 | | |
| **# of correctly classified named entities** | 170 | 180 | 194 |

Table 2.9: Evaluation of the NERC for Wiki-dictionary on the manually marked-up set

The highest accuracy, precision, recall and F1-score were achieved for the named entity label PER (see Table 2.10). The best results are for the version of the Wiki-dictionary with the labels translated from Russian mostlikely due to the better choice of the seed categories. However, as expected, precision is higher for the version of the Wiki-dictionary in which the intersection of the labels was chosen.

| | **intersection of En and Ru NE labels** | **NE labels translated from En** | **NE labels translated from Ru** |
|---|---|---|---|
| True positives | 61 | 62 | 64 |
| True negatives | 227 | 222 | 225 |
| False positives | 0 | 5 | 2 |
| False negatives | 12 | 11 | 9 |
| Accuracy | 96% | 94.67% | 96.33% |
| Precision | 100% | 92.54% | 96.97% |
| Recall | 83.56% | 84.93% | 87.67% |
| F1-score | 91.04% | 88.57% | 92.09% |

Table 2.10: Evaluation of the named entity label PER for Wiki-dictionary on the manually marked-up set

For the label LOC the accuracy, precision, recall and F1-score are much lower

than for the label PER and the statistics show that the dictionary with translated English labels has the worse results (see Table 2.11).

| | **intersection of En and Ru NE labels** | **NE labels translated from En** | **NE labels translated from Ru** |
|---|---|---|---|
| True positives | 61 | 62 | 74 |
| True negatives | 197 | 183 | 185 |
| False positives | 15 | 29 | 27 |
| False negatives | 27 | 26 | 14 |
| Accuracy | 86% | 81.67% | 86.33% |
| Precision | 80.26% | 68.13% | 73.27% |
| Recall | 69.32% | 70.45% | 84.09% |
| F1-score | 74.39% | 69.27% | 78.3% |

Table 2.11: Evaluation of the named entity label LOC for Wiki-dictionary on the manually marked-up set

The classes ORG and MISC are harder to identify than classes PER and LOC because the classes PER and LOC are well-represented by the hierarchy of categories in Wikipedia while the categories from Wikipedia that contain entities of the classes ORG and MISC are less structured. The precision and recall for these two classes are low and it is not obvious which version of Wiki-dictionary has better classification for these classes (see Table 2.12 and Table 2.13).

| | **intersection of En and Ru NE labels** | **NE labels translated from En** | **NE labels translated from Ru** |
|---|---|---|---|
| True positives | 2 | 4 | 3 |
| True negatives | 285 | 268 | 280 |
| False positives | 7 | 24 | 12 |
| False negatives | 6 | 4 | 5 |
| Accuracy | 95.67% | 90.67% | 94.33% |
| Precision | 22.22% | 14.29% | 20% |
| Recall | 25% | 50% | 37.5% |
| F1-score | 23.53% | 22.22% | 26.09% |

Table 2.12: Evaluation of the named entity label ORG for Wiki-dictionary on the manually marked-up set

The sample set of entries from the dictionary which was evaluated as correct is presented in the appendix A.

## 2.4 Dictionary Development for Other Language Pairs

Our implementation of the dictionary development can be easily applied to other language pairs. The heuristics that we used for the named entity recognition

| | **intersection of En and Ru NE labels** | **NE labels translated from En** | **NE labels translated from Ru** |
|---|---|---|---|
| True positives | 24 | 30 | 31 |
| True negatives | 236 | 230 | 213 |
| False positives | 17 | 23 | 40 |
| False negatives | 23 | 17 | 16 |
| Accuracy | 86.67% | 86.67% | 81.33% |
| Precision | 58.54% | 56.6% | 43.66% |
| Recall | 51.06% | 63.83% | 65.96% |
| F1-score | 54.55% | 60% | 52.54% |

Table 2.13: Evaluation of the named entity label MISC for Wiki-dictionary on the manually marked-up set

work only for English language so for language pairs that don't contain English a different method of named entity recognition should be chosen. However, the algorithm for named entity classification can work with any language pair.

We have built the English-German dictionary of the type I+R. To automatize the process of selection of the seed categories for the named entity classification, we collected the translations of the seed categories chosen for English from the Wikipedia. After running the NEC algorithm, we applied corrections based on the comments in brackets of the English titles.

In order to create the evaluation set for NERC, we took the manually marked up set of English-Russian entries and searched for the English entity page titles from this set in the English-German Wiki-dictionary. For the entries of English-German Wiki-dictionary that contained the titles of English entity pages we applied the same labels that the entries with these titles of English entity pages had in the English-Russian manually marked up set.

| | **intersection of English and German NE labels** | **NE labels translated from English** | **NE labels translated from German** |
|---|---|---|---|
| **# of entries in the sample** | 173 | | |
| **# of correctly classified named entities** | 84 | 99 | 99 |

Table 2.14: Evaluation of the NERC for English-German Wiki-dictionary on the evaluation set

The version of the dictionary with the labels translated from the results of classification of the German named entities was of best quality for classes PER (accuracy: 95.38%, precision: 93.62%, recall: 89.8%, F1-score: 91.67%) and MISC (accuracy: 86.7%, precision: 44.44%, recall: 60%, 51.06%) while the version of the dictionary with the labels translated from the results of the classification of the English named entities was of best quality for the class LOC (accuracy:

80.92%, precision: 63.58%, recall: 65.96%, F1-score: 65.26%) and had the highest accuracy for the class ORG (accuracy: 88.44%, precision: 5.88%, recall: 20%, F1-score: 9%).

The sample set of entries from the dictionary which was evaluated as correct is presented in the appendix B.

## 2.5 Dictionary evaluation

The evaluation of the dictionary generated from Wikipedia can be done manually and it can involve usage of the parallel corpora and/or machine-readable dictionary.

For example, in [9] the evaluation is done by comparison of the translations obtained from Wikipedia with the translations extracted from parallel corpus. The 200 domain terms for evaluation were selected from the parallel corpus by choosing 100 high frequency terms automatically and by choosing 100 low frequency words manually, so that 50 of the low frequency terms are included in the dictionary EDICT (Japanese-English Dictionary file)[15]. The authors calculated precision and recall for the translations extracted from the parallel-corpus and for the translations extracted from Wikipedia and compared the results. The precision and recall metrics were defined as following in their work:

$$\text{precision} = \frac{\text{Extracted correct translations}}{\text{All extracted translations}}$$

$$\text{recall} = \frac{\text{Extracted correct translations}}{\text{Correct translations in EDICT}}$$

For the evaluation the authors of [34] checked if the automatically collected translation pairs occurred in the parallel corpus of the debates of the European Parliament. They did the lemmatization first, and then counted occurrences in the corpus of a translation pair and the occurrence in the corpus of each lexical item of the translation pair. They applied the log likelihood ratio test to decide whether to keep or discard the translation pair.

### 2.5.1 Statistics in Mueller's dictionary

We evaluated if Wikipedia dictionaries cover the vocabulary of the unidirectional English-Russian dictionary by V. K. Mueller. We had machine readable version of the Mueller dictionary[16] in four plain text files: abbreviations (2204 entries), geographical names (1282 entries), names (630 entries) and base dictionary (50695 entries).

We removed comments in the brackets like

*Anna Karenina (1967 film)*
**Анна Каренина (фильм, 1967)**
*Transliteration: anna karenina (fil'm, 1967)*

[15] `http://www.csse.monash.edu.au/~jwb/edict.html`
[16] `http://sourceforge.net/projects/mueller-dict/files/dict/`

from the Wiki-dictionaries for the evaluation purpose because Mueller's dictionary English translations don't contain any comments in brackets.

The entries of the Mueller's dictionary are well-structured. The words in the base dictionary are supplied with their phonetic translation; grammatical (such as part of speech) and additional information (such as indication that the word is a special term) is provided in the form of the abbreviation that starts with the underscore symbol (e. g. "_refl.", "_pass.").

Roman figures mark homonyms, Arabic numerals with the dot mark different parts of speech, Arabic numerals with the bracket mark different meanings of the word, Russian letters with the bracket mark different meanings of the verbs with prepositions and phraseological expressions. Russian synonyms are separated by the comma and by the semicolon. We used this structure of the dictionary entry to calculate how many English-Russian pairs we can collect from the Mueller's dictionary.

| | **Mueller Base** | **Mueller Abbr.** | **Mueller Names** | **Mueller Geogr.** |
|---|---|---|---|---|
| # of entries | 50,695 | 2,204 | 630 | 1,282 |
| # of EN-RU pairs | 169,106 | 5,430 (2,850) | 933 | 1324 |

Table 2.15: Size of Mueller's Dictionaries

We managed to collect 2850 English-Russian pairs from the abbreviations file, 1324 pairs from the file with geographical names, 933 pairs from the names file and 169106 pairs from the base dictionary file (Table 2.15). We computed only the number of distinct lowercased English-Russian pairs that can be extracted from Mueller's dictionary because we did comparisons with the terms from Wiki-dictionary in lower case. However, our estimations of the number of Russian synonyms in the Mueller's dictionary are not 100% precise because of the ambiguity of the markup of the dictionary entries. For example, comma is sometimes part of the definition, e. g.

```
Home Counties
['hэum'kauntIz] _n.  _pl.  графства, окружающие Лондон
```
*Transliteration: grafstva, okružauščie london*

and sometimes it servers to separate the synonyms, e. g.

```
preciosity
[,preSI'OsItI] _n.  изысканность, утончённость, изощрённость
(языка, стиля)
```
*Transliteration: izyskannost, utončënnost, izoščrënnost (jazyka, stilja)*

We also didn't try to extract and compute the number of the English expressions and their translations given in the text of some entries, e. g.

```
a precious friend you have been!
_ирон.  хорош друг!
```
*Transliteration: _iron. xoroš drug!*

because we compared English terms from Wiki-dictionary only with the title English term in Mueller's dictionary. We didn't try to lookup English terms from Wiki-dictionary inside the body of the Mueller's entries (with the exception of the case when we are checking the extensions of the abbreviations).

Each article of the abbreviations file contains the possible English expansions of the abbreviation and alternative names mentioned in the brackets, for example:

```
A.T.
(тж.  A/T) American Terms американские технические условия
```
*Transliteration: (tž. A/T) American Terms amerikanskie texničeskie uslovija*

We wrote additional script that takes these properties of the entries of the abbreviations file and collected in total 5403 English-Russian pairs from this file using this approach.

We expected high coverage of abbreviations, geographical names and names because most of the entries in the Wiki-dictionaries are named entities. We couldn't expect high coverage of base dictionary, because most of the entries in the Wiki-dictionaries are nouns and phrases. However, the result numbers even for named entities are much lower than our expectations.

The algorithm worked the following way. It searched for the exact match of the lowercased English word from Mueller's dictionary in the Wiki-dictionary, e.g. we take a record

```
Czechoslovakia
_ист.  Чехословакия
```
*Transliteration: _ist. čexoslovakija*

from the Mueller's dictionary and search for the word 'czechoslovakia' in the Wiki-dictionary. If the entry of the Wiki-dictionary with such word is found, we collect all the Russian translations from the Wiki-dictionary. In our example the corresponding Wiki-dictionary record would be (the entry is shortened):

*Czechoslovakia | Federation of Czechoslovakia | Czechoslowakia | Czechaslavakia | CSFR*
**Чехословакия | Чехословацкая Социалистическая Республика | Чешско-Словацкая Социалистическая Республика | Чешско-Словацкая Федеративная Республика | ЧСФР**
*Transliteration: čexoslovakija | čexoslovackaja socialističeskaja respublika | češsko-slovackaja socialističeskaja respublika | češsko-slovackaja federativnaja respublika | čsfr*

We concatenate all the lines of the translation part in the Mueller's dictionary in one line and for each translation from the Wiki-dictionary we check if it occurs as a substring in Mueller's dictionary translation.

The reason why we concatenate all the lines of the translation part from Mueller in one line and search the Wiki-dictionary translations as a substring, is that translations in the Mueller's dictionary are quite often descriptive, e.g.

Zen
[zen] _n. буддийская секта ‘‘дзэн’’ (в Японии),
проповедующая созерцание и интуицию как основу прозрения
*Transliteration: buddijskaja sekta “dzèn” (v japonii),*
*propovedujuščaja sozercanie i intuiciju kak osnovu prozrenija*

If we would search only exact matches, we would get the result, that the Wiki-dictionary doesn't contain this pair. However, the Wiki-dictionary contains the record:

*Zen | Zen Buddhism | Ch'an | Zen buddhism | Zen Centers | Zen Glossary | Cha'an | Zen teachers | Zen monk | Zen Buddhist | Chan Buddhism | Chán | Chaan | Southern Chan | Yuanwu | Thien Buddhism | Thien | Thien buddhism | Vietnamese Zen | Viewing the phrase | Chinese Zen | Japanese Zen | Miyun Yuanwu | Zen Teacher | Ch'an Buddhism | Chán Buddhism | Zen+Buddhism | Zen Teacher/Zen Master | Zenshu | Zen Buddhists | Zen Philosophy | Zen philosophy*
**Дзэн | Дзен | Дзен-буддизм | Дзэн-буддизм | Цветочная проповедь Будды**
*Transliteration: dzèn | dzen | dzen-buddizm | dzèn-buddizm | cvetočnaja propoved buddy*

Using our approach of searching Russian translations as a substring in the Mueller's dictionary translations, we would detect this pair.

Many names were not found in the Wiki-dictionary, because the Mueller's dictionary contains the list of the most common first names, while Wikipedia titles usually contain the full names of the famous people and characters. For example, we don't find the name “Antoinette” in the Wiki-dictionary even though it contains such names as “Aimée Antoinette Camus”, “Antoine and Antoinette”, “Jean Baptiste Pierre Antoinette De Monet De Lamarck”, “Marie Antoinette (2006 film)”, “Marie Antoinette”, “Francoise Marie Antoinette Saucerotte”, “Antoinette Frissell” and others. To improve the recall for names, for each phrase pair from the Wiki-dictionary entry we compared the number of tokens and if the number of tokens was equal, we split the phrases into individual tokens.

We computed the recall of the Wiki-dictionaries with respect to the Mueller's dictionary; the diagram is shown on the Figure 2.5; the exact numbers are presented in the Table 2.16. For abbreviations and names we show improved statistics (achieved by looking-up abbreviation expansions and looking-up names in the dictionary with phrases split into independent tokens) and in the brackets we also provide original statistics.

The highest recall we obtained is, as expected, for the geographical names, however the value is lower than we expected: 82.18%. Original value for the names was low – 30.55%, but by splitting the names into tokens we improved recall up to 75.88%. Surprisingly, the highest recall we could get for the abbreviations, even taking the English expansions of the abbreviations into the account, is only 22.64%. The best recall we could obtain for the base dictionary is only 7.42%.

The main reason of rather high number of the unmatched named entities from Mueller's dictionaries is the fact that the titles of the Wikipedia articles

| Dictionary type | Mueller Base | Mueller Abbr. | Mueller Names | Mueller Geogr. |
|---|---|---|---|---|
| I (EN-DE-RU) | 9470 | 1067 (362) | 676 (157) | 1051 |
| I+R (EN-DE-RU) | 9557 | 1081 (367) | 687 (162) | 1055 |
| I | 11227 | 1154 (388) | 702 (248) | 1079 |
| I+R | 12240 | 1202 (401) | 708 (280) | 1087 |
| I+R+T | 12552 | 1223 (405) | 708 (285) | 1088 |

Table 2.16: Statistics of Wiki-dictionary with redirects on Mueller Dictionaries. The Wiki-dictionary types are defined in section 2.2.1.

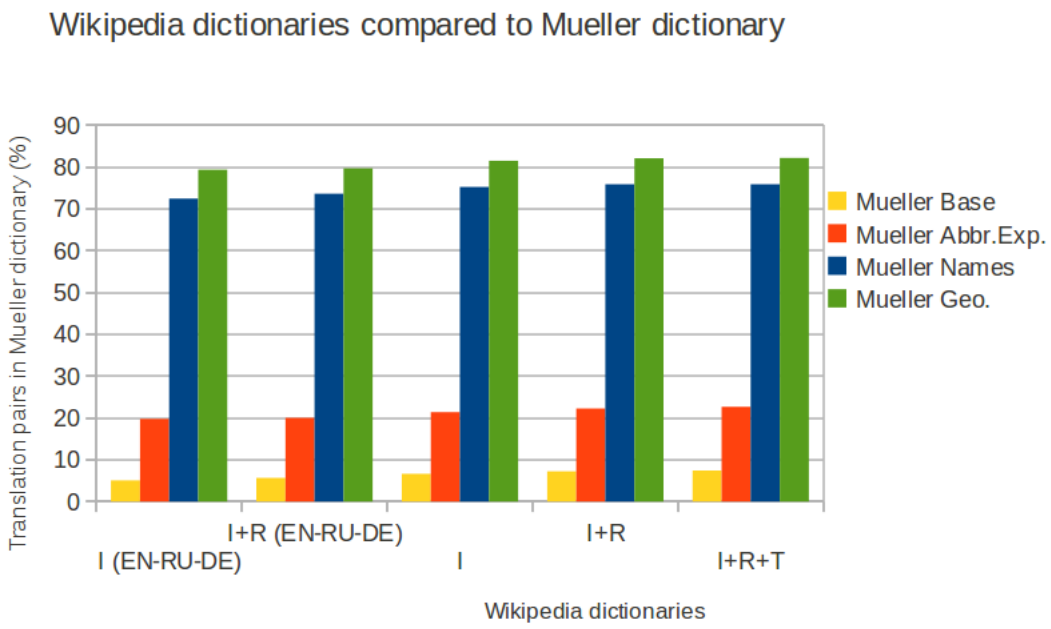



Figure 2.5: Wiki-dictionaries compared to Mueller dictionary; measure: RECALL

often contain additional details. For example, the entry

`Arkansas City`
`г. Арканзас-Сити`
*Transliteration: g. arkanzas-siti*

from the Mueller's dictionary cannot be matched with the entry from the Wiki-dictionary

*Arkansas City, Arkansas*
**Арканзас-Сити**
*Transliteration: arkanzas-siti*

as well as the entry

Amazon
р. Амазонка
*Transliteration: r. amazonka*

from the Mueller's dictionary cannot be matched with the entry from the Wiki-dictionary

*Amazon River*
**Амазонка**
*Transliteration: amazonka*

One of the other reasons of low result numbers for named entities is that we search for the exact match of the English term in the Mueller's dictionary and in the Wiki-dictionary. This means, that terms that differ only by word order or punctuation are considered as different records. Another reason is spelling: the name of the capital of Ghana is spelt "Akkra" in the Mueller's dictionary and is spelt "Accra" in the Wiki-dictionary, the name of the mount in Caucasus is spelt "Elbruz" in the Mueller's dictionary and is spelt "Elbrus" in the Wiki-dictionary[17]. Moreover, there are no related Russian articles for some English articles in Wikipedia and there are named entities among them.

The coverage of the base terminology of the Mueller's dictionary is very low because the Mueller's dictionary contains all parts of speech: nouns, verbs, adjectives, adverbs etc. while we expect that the Wiki-dictionary contains nouns and noun phrases. Table 2.17 shows how many pairs that occur in both Mueller's base dictionary and in Wiki-dictionaries are marked as nouns in the Mueller's base dictionary.

| | **I (EN-DE-RU)** | **I+R (EN-DE-RU)** | **I** | **I+R** | **I+R+T** |
|---|---|---|---|---|---|
| Total # of pairs found in Mueller's Base | 9,470 | 9,557 | 11,227 | 12,240 | 12,552 |
| Nouns in Mueller's Base | 9,211 | 9,295 | 10,836 | 11,799 | 12,101 |

Table 2.17: Statistics of phrases from Wiki-dictionary that have noun flag _n. in Mueller's Base Dictionary

Most of the other phrases have flag "_a." or "_v." in Mueller's base dictionary.

The flag "_n." occurs in Mueller's base dictionary 33,664 times and usually it is used to mark not a single word, but a group of possible translations. Even though most of the phrases of Wiki-dictionaries found in Mueller's base dictionary indeed have noun flag, it is obvious from the data we have, that Wiki-dictionaries don't cover one third of the nouns in Mueller's base dictionary. If we suppose that each noun flag in Mueller's dictionary marks a group of three nouns, we will

[17]Redirect page "Elbruz" is present in the current online version of Wikipedia.

have that I+R+T Wiki-dictionary covers only 11.98% of the nouns in Mueller's base dictionary.

The recall of the geographical names was lower than we expected, so we checked if the results are better or not for, for example, just English part of the Wiki-dictionary. Table 2.18 demonstrates that the numbers are a bit higher if we compute only the occurrence of English phrases, ignoring the Russian part of the Wiki-dictionary (compare with Table 2.16). We have the same picture if we will ignore the synonyms in the Wiki-dictionary (compare Table 2.18 with Table 2.19).

| **Dictionary type** | **Mueller Geogr.** | |
|---|---|---|
| | Without synonyms | With synonyms |
| I (EN-DE-RU) | 883 | 1135 |
| I+R (EN-DE-RU) | 885 | 1140 |
| I | 933 | 1166 |
| I+R | 942 | 1172 |
| I+R+T | 942 | 1172 |

Table 2.18: Statistics of the English part of the Wiki-dictionary with redirects and of the Wiki-dictionary with redirects on Mueller Geogr. Dictionary

Evaluations against the Mueller's dictionary also allowed us to compare the Wiki-dictionaries. We can see that relaxing of the condition of the intersection of English-Russian pairs with German, gave some improvement of the recall of geographical names: from 79.68% to 81.5%. As we expected, the recall of the Wiki-dictionaries increased when we relaxed the conditions and added extra records. The I+R+T dictionary had the highest recall on all the Mueller's dictionary files.

The statistics on Mueller's dictionary provides some evidence about the quality of the synonyms that we got from the titles of the redirect pages. The diagram on Figure 2.6 and Table 2.19 show the drop of the recall if we ignore synonyms in our evaluations.

| **Dictionary type** | **Mueller Base** | **Mueller Abbr.** | **Mueller Names** | **Mueller Geogr.** |
|---|---|---|---|---|
| I (EN-DE-RU) | 5665 | 497 (20) | 355 (115) | 747 |
| I+R (EN-DE-RU) | 5700 | 499 (20) | 357 (116) | 749 |
| I | 7024 | 549 (28) | 444 (204) | 782 |
| I+R | 7741 | 574 (30) | 461 (233) | 790 |
| I+R+T | 7910 | 585 (30) | 463 (240) | 791 |

Table 2.19: Statistics of Wiki-dictionary without synonyms on Mueller Dictionaries

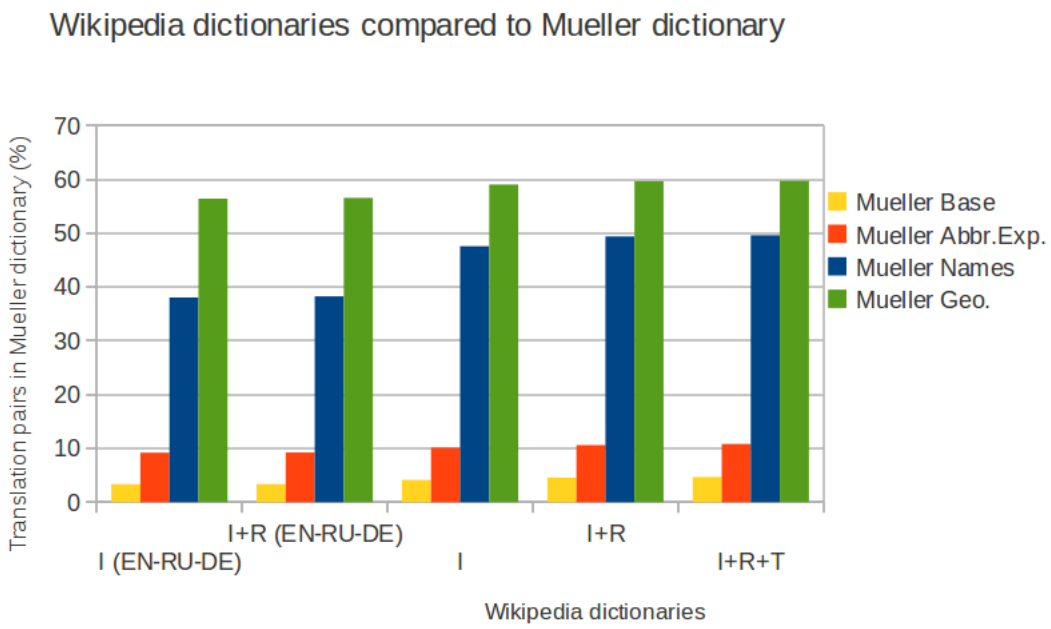



Figure 2.6: Wiki-dictionaries without synonyms compared to Mueller dictionary; measure: RECALL

### 2.5.2 Corpus Statistics

We collected the statistics of the occurrences of the translation pairs from the Wiki-dictionaries in the corpus. Each dictionary type we evaluated by word forms (using tokenized version of the dictionary developed in the section 2.2.3) and by normal forms (using tokenized normalized version of the dictionary developed in the section 2.2.4 and normalized version of the corpus data). These statistics were interesting for two reasons: first, we could evaluate the frequency of dictionary phrases in the corpus data and second, we had a goal to do pre-evaluation of the corpus to find out whether we could use it for the machine translation experiments with the dictionary.

The statistics was collected from the English-Russian parts of the Czech-English-Russian corpus which is a multilingual parallel corpus of texts in Czech, Russian and English languages [16],[21]. The corpus data was obtained from The Project Syndicate[18] which contains a huge collection of news articles and commentaries. We will call the corpus UMC (abbreviation from ÚFAL Multilingual Corpus). The texts in the UMC corpus are aligned only at sentence level; the data is tokenized and is available in the plain text format. The information about the size of the training data is presented in the Table 2.20

We didn't remove comments in brackets from the translation pairs of the Wiki-dictionaries, e. g.:

*Saints Peter and Paul Cathedral (Kazan)*
**Петропавловский собор (Казань)**
*Transliteration: petropavlovskij sobor (kazan')*

[18] `http://www.project-syndicate.org/`

| | Training set | | Devel. set | | Test set | |
|---|---|---|---|---|---|---|
| | **En** | **Ru** | **En** | **Ru** | **En** | **Ru** |
| **alignment pairs** | 86,604 | | 750 | | 1,000 | |
| **tokens** | 2,062,861 | 1,885,812 | 14,907 | 13,789 | 20,358 | 19,019 |
| **vocabulary size** | 42,045 | 113,747 | 3,639 | 5,409 | 4,728 | 7,001 |
| **average sentence length** | 23.82 | 21.78 | 19.88 | 18.39 | 20.36 | 19.02 |
| **standard dev. of sent. length** | 12.82 | 12.15 | 9.51 | 9.25 | 9.67 | 9.43 |

Table 2.20: UMC corpus statistics

We left the comments because we first performed experiments and evaluation on machine translation in chapter 3 with the dictionary with comments in brackets.

The details are summarized in Table 2.21, Table 2.22, Table 2.23. The tables show in how many sentences how many translation pairs from the dictionary occur. Statistics prove that English-Russian pairs from Wikipedia dictionary are present in the corpus. For example for the training set we see that 31 sentence contain 14 translation pairs from I (EN-DE-RU) dictionary, while 159 sentences contain 14 normalized translation pairs from the same dictionary.

Approximately 28% of the non-normalized training set doesn't contain any translation pair from the Wiki-dictionaries while approximately 24.7% of the the non-normalized training set contains exactly one translation pair from the Wiki-dictionaries. These statistics are lower for the normal forms: less sentences from the normalized training set contain zero and one normalized translation pair, but more sentences contain more translation pairs.

The distribution is similar for the development and test sets, though the actual values are, obviously, much lower due to the fact that the sets are much smaller than the training set. The development and test sets have comparable size therefore the numbers are quite similar for them.

| | NUMBER OF SENTENCES | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (COUNTS BY WORD FORMS) | | | | | (COUNTS BY NORMAL FORMS) | | | | |
| # of pairs | I (EN-DE-RU) | I +R (EN-DE-RU) | I | I +R | I +R +T | I (EN-DE-RU) | I +R (EN-DE-RU) | I | I +R | I +R +T |
| 51 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 46 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| 43 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 40 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 38 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 36 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 3 |
| 35 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 2 |

| | NUMBER OF SENTENCES | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (COUNTS BY WORD FORMS) | | | | | (COUNTS BY NORMAL FORMS) | | | | |
| # of pairs | I (EN-DE-RU) | I +R (EN-DE-RU) | I | I +R | I +R +T | I (EN-DE-RU) | I +R (EN-DE-RU) | I | I +R | I +R +T |
| 33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 |
| 32 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 2 | 1 |
| 31 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 3 | 3 |
| 30 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 3 | 3 | 4 |
| 29 | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 2 |
| 28 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 3 | 2 | 3 |
| 27 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 6 | 8 | 9 |
| 26 | 1 | 1 | 0 | 0 | 0 | 2 | 2 | 6 | 10 | 10 |
| 25 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 12 | 9 | 11 |
| 24 | 1 | 1 | 1 | 1 | 2 | 8 | 8 | 9 | 16 | 17 |
| 23 | 2 | 2 | 2 | 2 | 1 | 3 | 3 | 14 | 19 | 37 |
| 22 | 0 | 0 | 0 | 2 | 2 | 6 | 7 | 39 | 44 | 31 |
| 21 | 1 | 1 | 2 | 3 | 4 | 6 | 5 | 37 | 34 | 37 |
| 20 | 3 | 3 | 5 | 5 | 4 | 21 | 21 | 35 | 50 | 53 |
| 19 | 5 | 5 | 6 | 7 | 7 | 16 | 16 | 62 | 79 | 90 |
| 18 | 5 | 5 | 11 | 7 | 7 | 28 | 28 | 86 | 109 | 127 |
| 17 | 3 | 3 | 8 | 16 | 17 | 41 | 41 | 133 | 151 | 168 |
| 16 | 8 | 8 | 18 | 23 | 27 | 66 | 67 | 184 | 238 | 235 |
| 15 | 9 | 10 | 19 | 34 | 36 | 107 | 108 | 254 | 319 | 337 |
| 14 | 31 | 30 | 51 | 45 | 55 | 159 | 160 | 350 | 424 | 485 |
| 13 | 42 | 43 | 69 | 87 | 83 | 219 | 224 | 540 | 596 | 640 |
| 12 | 54 | 55 | 109 | 121 | 127 | 345 | 349 | 736 | 862 | 888 |
| 11 | 101 | 100 | 173 | 220 | 237 | 500 | 495 | 1022 | 1154 | 1242 |
| 10 | 167 | 170 | 281 | 342 | 365 | 789 | 796 | 1473 | 1667 | 1708 |
| 9 | 261 | 262 | 501 | 593 | 635 | 1135 | 1145 | 2002 | 2238 | 2333 |
| 8 | 583 | 587 | 839 | 974 | 1048 | 1725 | 1737 | 2905 | 3164 | 3293 |
| 7 | 843 | 846 | 1363 | 1549 | 1575 | 2569 | 2575 | 3946 | 4237 | 4326 |
| 6 | 1515 | 1525 | 2335 | 2590 | 2724 | 3910 | 3919 | 5345 | 5729 | 5842 |
| 5 | 2698 | 2708 | 3908 | 4297 | 4410 | 5881 | 5890 | 7332 | 7571 | 7653 |
| 4 | 4914 | 4943 | 6379 | 6869 | 7081 | 8387 | 8384 | 9360 | 9492 | 9539 |
| 3 | 8658 | 8657 | 10435 | 10782 | 10851 | 11632 | 11630 | 11703 | 11537 | 11484 |
| 2 | 14841 | 14881 | 15758 | 15939 | 16035 | 15311 | 15335 | 13423 | 12957 | 12708 |
| 1 | 22386 | 22367 | 21402 | 20575 | 20269 | 17204 | 17168 | 13646 | 12742 | 12444 |
| 0 | 29470 | 29389 | 22925 | 21516 | 20997 | 16521 | 16478 | 11923 | 11125 | 10833 |

Table 2.21: Pair statistics of Wiki-dictionaries on the UMC train set

The statistics is better if we remove the comments in brackets. It is presented in the appendix C. However, sometimes comments are separated by the comma

| | NUMBER OF SENTENCES | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (COUNTS BY WORD FORMS) | | | | | (COUNTS BY NORMAL FORMS) | | | | |
| # of pairs | I (EN-DE-RU) | I+R (EN-DE-RU) | I | I+R | I +R +T | I (EN-DE-RU) | I+R (EN-DE-RU) | I | I+R | I +R +T |
| 23 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 21 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 16 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| 15 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 2 | 1 | 0 |
| 14 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 2 | 3 |
| 13 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 1 |
| 12 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 4 | 6 | 6 |
| 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 6 | 7 | 8 |
| 9 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 6 | 10 | 7 |
| 8 | 1 | 1 | 1 | 2 | 2 | 6 | 6 | 16 | 16 | 17 |
| 7 | 5 | 5 | 6 | 7 | 7 | 9 | 9 | 21 | 20 | 23 |
| 6 | 7 | 7 | 12 | 13 | 14 | 21 | 21 | 40 | 46 | 50 |
| 5 | 14 | 14 | 18 | 21 | 23 | 44 | 44 | 61 | 67 | 68 |
| 4 | 24 | 25 | 49 | 57 | 55 | 65 | 66 | 73 | 76 | 72 |
| 3 | 78 | 77 | 91 | 95 | 98 | 117 | 116 | 125 | 126 | 125 |
| 2 | 136 | 137 | 141 | 143 | 143 | 147 | 149 | 131 | 130 | 130 |
| 1 | 202 | 201 | 207 | 200 | 199 | 158 | 156 | 142 | 132 | 130 |
| 0 | 281 | 281 | 223 | 210 | 207 | 173 | 173 | 119 | 108 | 106 |

Table 2.22: Pair statistics of Wiki-dictionaries on the UMC development set

which causes ambiguity because some translations without comments contain commas. We didn't try to disambiguate such cases.

| | NUMBER OF SENTENCES | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (COUNTS BY WORD FORMS) | | | | | (COUNTS BY NORMAL FORMS) | | | | |
| # of pairs | I (EN-DE-RU) | I+R (EN-DE-RU) | I | I+R | I +R +T | I (EN-DE-RU) | I+R (EN-DE-RU) | I | I+R | I +R +T |
| 21 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 4 | 3 |
| 12 | 0 | 0 | 0 | 0 | 1 | 2 | 2 | 8 | 6 | 7 |
| 11 | 0 | 0 | 0 | 1 | 1 | 6 | 6 | 8 | 11 | 10 |
| 10 | 0 | 0 | 1 | 3 | 2 | 5 | 5 | 11 | 10 | 15 |
| 9 | 3 | 3 | 7 | 5 | 5 | 9 | 11 | 18 | 27 | 24 |
| 8 | 3 | 3 | 5 | 6 | 6 | 10 | 8 | 21 | 20 | 22 |
| 7 | 9 | 9 | 10 | 12 | 17 | 21 | 21 | 31 | 39 | 42 |
| 6 | 9 | 9 | 19 | 26 | 23 | 40 | 40 | 64 | 72 | 75 |
| 5 | 24 | 25 | 30 | 28 | 29 | 62 | 63 | 83 | 79 | 77 |
| 4 | 37 | 36 | 44 | 55 | 57 | 92 | 92 | 109 | 117 | 122 |
| 3 | 108 | 109 | 139 | 147 | 149 | 161 | 161 | 150 | 149 | 146 |
| 2 | 185 | 185 | 202 | 200 | 204 | 193 | 193 | 176 | 165 | 162 |
| 1 | 262 | 263 | 268 | 259 | 258 | 194 | 193 | 173 | 162 | 160 |
| 0 | 360 | 358 | 275 | 258 | 248 | 204 | 204 | 143 | 136 | 131 |

Table 2.23: Pair statistics of Wiki-dictionaries on the UMC test set

# 3. Evaluation on Machine Translation System

## 3.1 Machine Translation: Background

### 3.1.1 Statistical Approach to Machine Translation

With the publication [2] statistical methods and ideas from information theory gained popularity in the field of machine translation.

We translate a foreign sentence $f$ to the sentence $e$ in another language. This process can be modeled using the noisy-channel model as shown on the Figure 3.1.


Figure 3.1: Noisy channel

We assume that the foreign speaker actually wanted to utter a sentence $e$ but everything got distorted in a noisy channel and out came a sentence $f$ in the foreign language. Then we can give the following formulation of our task: given a sentence $f$ we seek for the sentence $e$ from which the translator produced $f$. We choose such a sentence $e$ which is the most probable given $f$.

$$Pr(e|f) = \frac{Pr(e) * Pr(f|e)}{Pr(f)} \text{ (Bayes' theorem)}$$

$$argmax_e Pr(e|f) = argmax_e Pr(e) * Pr(f|e)$$

($Pr(f)$ doesn't depend on $e$)

$Pr(e)$ - the language model probability of $e$

$Pr(f|e)$ - the translation probability of $f$ given $e$

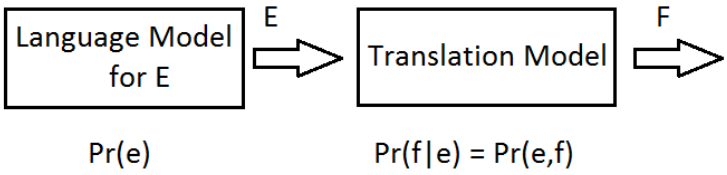



Figure 3.2: Language model and translation model

The parameters of language and translation models are estimated automatically from a large database of source-target sentence pairs using a statistical algorithm.

A decoder performs the actual translation. Given a sentence $f$, the decoder chooses such a sentence $e$ for which the probability $Pr(e|f)$ is maximal, the illustration is provided on Figure 3.3.

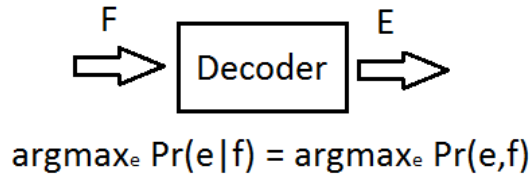



Figure 3.3: Decoder

### 3.1.2 Phrase-based Machine Translation

The authors of [20] showed that machine translation models based on phrases of up to three words outperform word-based models. The phrase translation model is based on the noisy channel model. The authors used Bayes rule:

$$argmax_e p(e|f) = argmax_e p(e) * p(f|e)$$

where $f$ is a sentence in a foreign language from which we translate and $e$ is a sentence in a language into which we translate.

During decoding the foreign sentence $f$ is segmented into a sequence of phrases and each foreign phrases is translated into an original language phrase. Phrases can be reordered.

The modern methods of learning phrase translation table from a parallel corpus start with a word alignment. Alignment is an object for indicating the corresponding words in a parallel text [4]. Word alignment task is often difficult or humans especially the cases of idiomatic expressions, free translations and missing function words. Often an alignment between two word strings includes such effects as reordering, omissions, insertions and word-to-phrase alignments [28]. A freely available toolkit GIZA++ is commonly used for word alignment task. GIZA++ is an implementation of the IBM Models. These models were developed for word-based machine translation where the translation relied on the translation of words in isolation.

- IBM Model 1 introduces lexical translation.
- IBM Model 2 adds absolute alignment model.
- IBM Model 3 adds fertility model.
- IBM Model 4 adds relative alignment model.
- IBM Model 5 fixes deficiency.

To overcome the weaknesses of IBM Models (such as the limitation that at most one English word can be aligned with each foreign word), the parallel corpus is aligned bidirectionally. Taking intersection or union of the two alignments we get high-precision or high-recall alignments correspondingly.

A bilingual dictionary can be used as an additional knowledge source for training of the alignment models. The parameters of the alignment models can be estimated by applying the EM algorithm. A dictionary is assumed to be a list of word strings $(e, f)$ where $e$ and $f$ can be single words or phrases.

One of such methods of integrating of the dictionary into EM algorithm, described in [3], requires adding every dictionary entry $(e, f)$ to the training corpus with an entry-specific count called effective multiplicity expressed as

$$\mu(e, f) = \frac{\lambda(e)p(f|e)}{1 - e^{-\lambda(e)p(f|e)}}$$

Results of experiments in [3] showed that the dictionary helps to improve the fertility probabilities for rare words.

Another method described in [27] suggests that effective multiplicity $\mu(e, f)$ of a dictionary entry should be set to a large number if the lexicon entry occurs in at least one of the sentence pairs of the bilingual corpus and to low value if it doesn't occur in the corpus. The approach helps to avoid a deterioration of the alignment as a result of a out-of-domain dictionary entries.

### 3.1.3 Machine Translation Evaluation

Evaluation of machine translation quality is not a trivial task. Human translators could be asked to judge the translation quality or the similarity between the machine translation system output and translations created by human translators could be computed automatically. One of the difficulties is that different human translators usually come up with different translations even for short sentences. Another aspect of the problem is that we are not only interested if translation is correct or wrong, but we would also like to evaluate how useful the output of machine translation system is for human-aided machine translation, e.g. how much effort is required to post-edit the results.

The main practical requirement for the evaluation metric is the low cost, measured in time and money. We would like to have a quick and cheap method to carry out evaluation. The metric should also be consistent which means that different evaluators should come to the same conclusions if they use the same metric (inter-annotator agreement) and the evaluation on one part of the test corpus should be consistent with the evaluation on the other part. And finally, evaluation metric should be correct.

#### 3.1.3.1 Manual Evaluation

The plausible evaluation method is to invite human translators who could evaluate by hand whether the translation is correct or wrong. In the situations when human translators who know both source and target languages are not available, the evaluation could be done by the evaluator with the knowledge of the target language only. The evaluation is usually done sentence by sentence as well as sentence in a context because such tasks as coreference resolution often require longer contexts.

The evaluation system that allows to judge if the translation is perfect or not is rarely being used because nowadays it is appropriate only for short sentences. A more common approach is to use a graded scale with the two criteria – fluency and adequacy.

Fluency estimates the fluency of English from grammatical and idiomatic points of view.

Adequacy estimates how close the meaning of the translation is to the input.

When the evaluation is carried out by several human evaluators, the normalization of the judgments is required because the scores assigned to the same sentence by different people usually fall in the wide range. The average of a set of judgments $\{x_1, \ldots, x_n\}$ is defined by the formula:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Judging fluency and adequacy is a tricky task because humans have the ability to fill the information gaps from their background knowledge. Having read the input sentence in the source language first, an evaluator may not notice that the output of the system doesn't cover the original meaning.

The easier task is to rank two or more systems against each other on a sentence-by-sentence basis in order to decide if the output of one system is indistinguishable, better or worse than the output of another system.

#### 3.1.3.2 Automatic evaluation

Each system translation could be compared to one or more reference translations (human translations). The higher the similarity score is, the more likely it is that the translation is correct. The common metrics are precision and recall. Precision computes how many words that a system generates are correct.

$$\text{precision} = \frac{\text{correct}}{\text{output-length}}$$

Recall metric computes how many of the words that system should generate are correct.

$$\text{recall} = \frac{\text{correct}}{\text{reference-length}}$$

Both metrics are equally important in machine translation (as we don't want to generate wrong words and we don't want the output to lack required words either) which is the motivation to use a harmonic mean of precision and recall - f-measure.

$$\text{F1-score} = \frac{\text{correct}}{(\text{output-length} + \text{reference-length})/2}$$

Position-independent error rate metric measures mismatches and punishes too long translations.

$$\text{PER} = 1 - \frac{\text{correct} - max(0, \text{output-length} - \text{reference-length})}{\text{reference-length}}$$

Word-error rate takes word order into consideration. This metric is based on Levenshtein distance, minimal number of editing steps – insertions, deletions and substitutions, - needed to match two sequences.

$$\text{WER} = \frac{\text{substitutions} + \text{insertions} + \text{deletions}}{\text{reference-length}}$$

Metric BLEU (bilingual evaluation understudy) is defined by the formula:

$$\text{BLEU-n} = \text{brevity-penalty} * exp \sum_{i=1}^{n} \lambda_i \log \text{precision}_i$$

$$\text{brevity-penalty} = min(1, \frac{\text{output-length}}{\text{reference-length}})$$

BLEU overcomes the drawback of precision metric – no penalty for dropping words – by introducing a brevity penalty parameter, which punishes too short sentences.

$n$ is the maximum order for n-grams to be matched, it is usually chosen to be equal to 4.

$\{\lambda_i\}$ are the weights for the different precisions. They are often all set to 1.

The simplified formula for BLEU is:

$$\text{BLEU-4} = min(1, \frac{\text{output-length}}{\text{reference-length}}) * \prod_{i=1}^{4} \text{precision}_i$$

BLEU scores are typically computed over the whole test set because on the sentence level we often have n-gram precisions that equal to 0 which leads to zero BLEU score.

It is possible to use multiple reference translations for evaluation with BLEU. If n-gram in the system output matches n-gram in any of the reference translations, it is considered as correct. When multiple-reference BLEU is used, the reference length is defined as the closest length to any of the reference translations.

Among the main limitations of BLEU are the following shortcomings:

- BLEU doesn't take into account that some words are more important than others, e.g. negation word 'not' and names are more important than determiners and punctuation.
- BLEU works only at n-gram level therefore it cannot check the overall grammatical coherence. Moreover, as explained in [21], BLEU cannot be reliable indeed when applied to languages with free word order because many n-grams won't match when words are swapped.
- BLEU score compares the word forms, so for the languages with the rich inflection, minor translation variations or errors can cause the unfair loss in BLEU score.
- The actual BLEU score is meaningless because the value depends on many factors such as the language pair, the domain, the number of reference translations and others.

## 3.2 Experiments with Moses

Nowadays several open-source toolkits for phrase-based statistical machine translation are available such as Moses[1] and Joshua[2]. We chose Moses for our experiments. This toolkit is a complete machine translation system for academic

[1] `http://www.statmt.org/moses/`

[2] `http://joshua.sourceforge.net/Joshua/Welcome.html`

research [19]. It consists of different modules for data pre-processing, language and translation models training and evaluating the resulting translations using the BLEU score. Moses uses external tools for some of the tasks, such as GIZA++ for word alignments, SRILM for language modeling and MERT for tuning the models using minimum error rate training. The core modules of Moses toolkit are shown on the Figure 3.4.

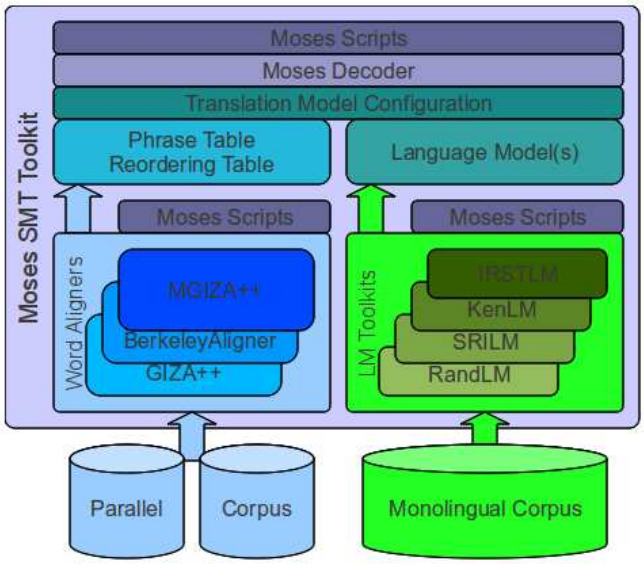



Figure 3.4: Basic components of Moses toolkit

Additionally, Moses has instruments for phrase-based translation with factors which allows to augment the surface forms of the words with different factors, such as part-of-speech tags or lemma. The core component of Moses is the decoder which replaced the popular phrase-based decoder Pharaoh[3]. The decoder in Moses is extended to process multiple, ambiguous hypotheses in the form of an input lattice which is approximated by confusion network.

We used freely available SRI language modeling toolkit[4] to build the language models. This software creates a model based on N-gram statistics from training data and computes the probability of a test corpus, conventionally expressed as the test set perplexity. For smoothing we chose modified Kneser-Ney discounting and interpolation.

The Kneser-Ney algorithm differs from other discounting techniques by the method of optimization of the lower-order distributions [32]. For example, in bigram case the unigram probability is proportional to the number of different words that it follows (instead of being proportional to the number of occurrences of a word).

The main difference between backoff and interpolated models is that in determining the probability of n-grams with nonzero counts, interpolated models use information from lower-order distributions while backoff models do not. But both backoff and interpolated models use lower-order distributions to assign probability to n-grams with zero counts.

[3] `http://www.isi.edu/licensed-sw/pharaoh/`

[4] `http://www.speech.sri.com/projects/srilm/`

The main knowledge source for the machine translation decoder are the translation tables. The decoder consults these tables to figure out how to translate input in one language into output in another language. This is a line from the phrase table from one of our experiments:

```
( хоккей ) ||| ( ice hockey ) ||| (0) (1,2) (3) ||| (0) (1) (1) (2)
||| 0.0526316 0.00526094 0.6 0.00368952 2.718
```

The columns of the table separated by the symbol |||.
The first column contains the phrase in the source language: "( хоккей )".
The second column contains the phrase in the target language: "( ice hockey )".
The third column outputs the word-to-word alignments between f (the phrase in the source Russian language in our case) and e (the phrase in the target English language in our case):

```
( → (
хоккей → ice hockey
) → )
```

The fourth column outputs the word-to-word alignments between e (the phrase in the target English language in our case) and f (the phrase in the source Russian language in our case) because we specified the option msd-bidirectional-fe for GIZA++:

```
( → (
ice → хоккей
hockey → хоккей
) → )
```

The fifth group shows five different phrase translation scores:

1. inverse phrase translation probability $\psi(f|e) = 0.0526316$;
2. inverse lexical weighting $lex(f|e) = 0.00526094$;
3. direct phrase translation probability $\psi(e|f) = 0.6$;
4. direct lexical weighting $lex(e|f) = 0.00368952$;
5. phrase penalty (always $exp(1) = 2.718$);

The default weights are computed during the training of the language model. However these weights are not guaranteed to be of high quality therefore tuning is needed, e. g. a procedure that would find the weights that minimize a given error measure, or, equivalently, maximize a given translation metric. The standalone open-source software for minimum error rate training (MERT, [1]) is implemented within the Moses toolkit. The two important components of the MERT are *the scorer*, that uses automatic metrics BLEU in PER (position independent error rate) to score given ranking of the n-best lists, and the *optimizer*, that performs the actual parameter optimization.

We used UMC corpus introduced in section 2.5.2 for the machine translation experiments. We trained and evaluated several translation models, that are presented with their BLEU and NIST scores in the Table 3.1. First, we performed several experiments without Wiki-dictionary and achieved the highest BLEU score 24.76 using additional data for training language model. Then we incorporated the dictionary into the training set: the dictionary was split into pairs of synonyms and appended to the end of the UMC corpus. Each synonym of the of the dictionary played role of the sentence of the corpus.

| | **BLEU score** | **NIST score** |
|---|---|---|
| 3-gram | 21.19 | 6.0225 |
| 4-gram | 21.42 | 6.0651 |
| 5-gram | 20.99 | 5.9800 |
| 4-gram + additional data for LM | 24.60 | 6.2988 |
| 5-gram + additional data for LM | 24.76 | 6.2742 |
| 3-gram + Wiki-dict. | 20.05 | 5.8498 |
| 4-gram + Wiki-dict. | 20.42 | 5.9314 |
| 5-gram + Wiki-dict. | 20.38 | 5.9198 |

Table 3.1: BLEU score for the trained models. "3-gram", "4-gram", "5-gram" describes the size of the language model, "additional data for LM" means the English monolingual data from the Europarl corpus (`http://www.statmt.org/europarl/`) was added to the UMC training set to train the language model for English, "Wiki-dict." means the Wiki-dictionary was incorporated into the UMC training set.

The Table 3.2 shows the individual and cumulative n-gram scoring for the two models (one trained without Wiki-dictionary and the other trained with Wiki-dictionary) that achieved the highest BLEU score without additional data for language model training.

| | **type of scoring** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|---|---|
| **w/t Wiki-dict** | **Individual N-gram scoring** | 0.55 | 0.27 | 0.15 | 0.09 | 0.05 | 0.03 | 0.02 | 0.01 |
| | **Cumulative N-gram scoring** | 0.55 | 0.39 | 0.28 | 0.21 | 0.16 | 0.13 | 0.10 | 0.08 |
| **with Wiki-dict** | **Individual N-gram scoring** | 0.54 | 0.26 | 0.14 | 0.08 | 0.05 | 0.03 | 0.02 | 0.01 |
| | **Cumulative N-gram scoring** | 0.54 | 0.38 | 0.27 | 0.20 | 0.15 | 0.12 | 0.09 | 0.07 |

Table 3.2: Individual and cumulative n-gram scoring for 4-gram models trained with and without Wiki-dictionary

### 3.2.1 OOV Words Analysis (UMC Test Set)

OOV (out-of-vocabulary) words are the words of the source language (Russian in our case) that the machine translation system didn't manage to translate into the target language (English in our case).

For the models that were trained just on the training set of the corpus the total number of different OOV words was 934 (988 in total, when we count repeated OOV words). When we integrated Wikipedia dictionary into the training set, the total number of different OOV words decreased down to 906 (945 in total if we count repeated OOV words).

But this doesn't mean that in both cases we have the same set of 906 OOV words. In fact, in both cases we have the same set of 699 OOV words. 235 words that were OOV with previous models are not OOV when using Wikipedia dictionary(approximately one third of these words are names of people and places). However, 207 words that were not OOV before, became OOV when we applied Wiki-dictionary.

Moreover, some of OOV words among those 235 were not translated with Wikipedia dictionary, but simply became missing words. For example,without Wikipedia we had the sentence

```
<seg id=414> for example , the czechs and slovaks already
отделились from each other , as well as many peoples yugoslavia .
</seg>
```

When Wikipedia was used, the word "отделились" ("separated"; transliteration: otdelilis) was missing in the translation:

```
<seg id=414> for example , the czechs and slovaks already with
each other , as well as many people in yugoslavia .  </seg>
```

Among the OOV words that were correctly translated using the Wikipedia dictionary are:

нападении (attack; transliteration: napadenii)
московская (moscow; transliteration: moskovskaja)
поединок (duel; transliteration: poedinok)
рефлекс (reflex; transliteration: refleks)
пенджаба (punjab; transliteration: pendžaba)
нагаленда (nagaland; transliteration: nagalenda)
ассама (assam; transliteration: assama)
мышьяк (arsenic; transliteration: myšjak)

But some of the OOV words were translated incorrectly after applying Wikipedia dictionary, e.g. патрон (transliteration: patron), was incorrectly translated as "cartridge" in the sentence

```
<seg id=21> syriac cartridge hezbollah , president bashar al
-асад ,one of the members of the `` axis of evil , '' also feels very
well .</seg>
```

One of the reasons is polysemy: the word “патрон” (transliteration: patron) has senses “patron” and “cartridge”.

Among the words that were correctly translated without Wikipedia dictionary but became OOV when Wikipedia dictionary was applied are:

драматическими (dramatic; transliteration: dramatičeskimi)
продолжительными (long; transliteration: prodolžitelnymi)
переговорный (negotiating; transliteration: peregovornyj)

Some words that became OOV after the dictionary was integrated into the training set, had been translated correctly before, such as “резней” (“slaughter”; transliteration: reznej), “мучительные” (“painful”; transliteration: mučitelnye), “ездить” (“travel”, transliteration: ezdit), “пережиток” (“relic”; transliteration: perežitok) though some had been translated incorrectly, such as “лишённые” (“excluded from”; transliteration: lišënnye) incorrectly translated as “disillusioned”.

The diagrams on the Figure 3.5 show the distribution of OOV along the UMC test set. Both diagrams have similar shape and most sentences have 2-3 OOV words for both language models. The number of OOV words tends to be a little higher along the whole test set for the language model trained without Wiki-dictionary.

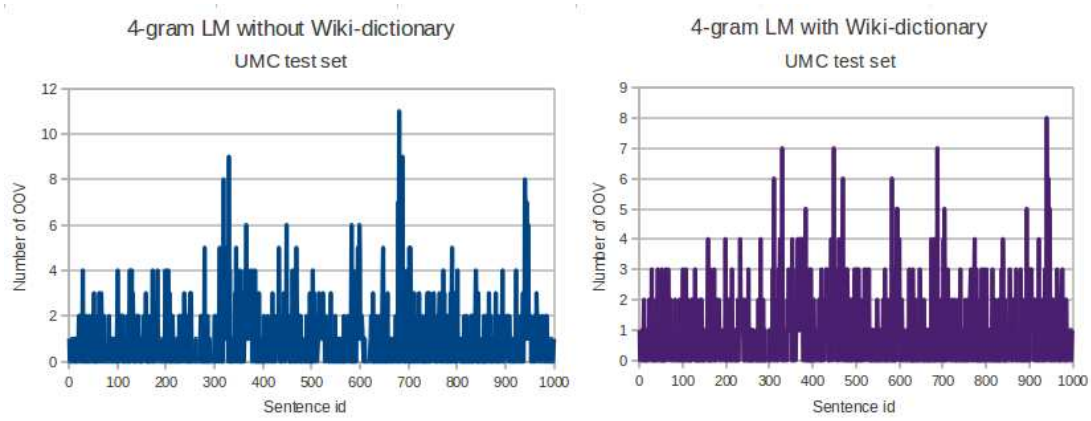



Figure 3.5: Distribution of OOV words on the UMC test set for the 4-gram model trained without Wiki-dictionary (on the left) and for the 4-gram model trained with Wiki-dictionary (on the right)

### 3.2.2 Paired Bootstrap Re-Sampling (UMC Test Set)

We used paired bootstrap re-sampling to estimate the statistical significance of the the difference in BLEU score between the model created with and without Wikipedia dictionary. As the difference between BLEU scores of the systems was small, we couldn’t be sure if we could trust automatic evaluation results that one system outperformed the other on the test set. Our question was if the difference in test scores was statistically significant.

The approach is described in [18]. We collected 1000 trial sets of the size 300 sentences from the original test set (which had the size of 1000 sentences) by random sampling with replacement. We computed BLEU score for both systems

in question on each of the 1000 trial sets and calculated how many times one system outperformed the other.

We compared the models that were created without additional corpus for language model training. The results are summarized in the Table 3.3. According to our evaluation, 3-gram model without Wikipedia dictionary is better than the model trained with Wikipedia dictionary with 98.5% statistical significance, 4-gram model is better with 96% statistical significance and 5-gram model is better with 87.1% statistical significance.

| **Model 1** | **Model 2** | **Statistical significance that model 1 outperforms model 2** |
|---|---|---|
| 3-gram | 3-gram + Wiki-dict. | 98.5% |
| 4-gram | 4-gram + Wiki-dict. | 96% |
| 5-gram | 5-gram + Wiki-dict. | 87.1% |

Table 3.3: The results of the paired boostrap re-sampling show the statistical significance of the fact that the models trained without Wiki-dictionary outperform the models trained with Wiki-dictionary

### 3.2.3 Manual Ranking (UMC Test Set)

We also performed manual evaluation of the 4-gram models by ranking the translations as described in section 3.2 of [6]. First, we collected all the translations that differ with diff command. It turned out to be almost the whole test set. We collected a sample of the size 100 sentences from those that differ the following way: as almost all the translations differ, we took one sentence from each ten, e.g. we took sentences number 2, 11, 20, 33, 47, 55 and so on.

In most of the cases one system was ranked higher than the other because of the better representation of the meaning of the original sentence. In many other cases the missing words and grammatical structure played the key role in the final decision. There were several pairs for which one translation was preferred against the other because of the vocabulary, as some synonyms suit particular contexts better than the other synonyms.

The final result is: for 55 sentences the model that didn't use Wikipedia dictionary was ranked higher, for 8 sentences both systems were ranked equally (equally bad/good translations) and for 37 sentences the system trained with Wikipedia dictionary was ranked higher. For the first 50 sentences from our trial set the model that didn't use Wikipedia dictionary was ranked higher for 25 sentences , both models were ranked equally for 5 sentences and the model trained with Wikipedia dictionary was ranked higher for 20 sentences. These results are summarized in the Table 3.4.

The overall results of the manual evaluation correspond to the automatic evaluation.

Typically, the more data is used for the translation model training the higher translation performance can be achieved. However, the significant amount of out-of-domain data added to the training set cause the drop of the translation quality [13]. In such a case a general translation model P(s|t), that was trained on in-domain and out-of-domain data, does not fit the topic or style of individual

| | Model w/t Wiki-dict is ranked first | Model with Wiki-dict is ranked first | Translations are equally bad/good |
|---|---|---|---|
| **ranking of first 50 sentences** | 25 | 20 | 5 |
| **ranking of 100 sentences** | 55 | 37 | 8 |

Table 3.4: The results of the manual ranking of the translations produced by the 4-gram language model trained without Wiki-dictionary and the 4-gram language model trained with Wiki-dictionary on the sample from the UMC test set

texts. For the ambiguous words the translation highly depends on the topic and context they are used in.

Evaluation in the section 2.5.2 the UMC training set contained a significant number of sentences that contained zero or only one word from the Wiki-dictionary. We believe that might mean that the domains of the Wiki-dictionary and the UMC corpus are quite different. We suppose that was the reason of the lower quality of the translation that we got from the model trained on the train set with the Wiki-dictionary incorporated in it.

### 3.2.4 Data Collection for the Wiki-set

To support or reject our hypothesis that the level of translation quality dropped because of the inappropriateness of the domain, we carried out another experiment. We collected three Wikipedia articles in the versions that were available on July 19, 2011. The articles were dedicated to three famous people: Madonna (singer), Freddie Mercury and Vladimir Lenin.

The text of the articles needed pre-processing. First, we converted MediaWiki text into plain text using the Java Wikipedia API (Bliki engine)[5] which is a parser library for converting Wikipedia wikitext notation to other formats. The class PlainTextConverter from this library can convert simple Mediawiki texts to plain text. Secondly, we removed that traces of template markup (e. g. {{cite web}} ) that still remained after removing Mediawiki markup. Thirdly, we split the text into sentences with the script split-sentences.perl written by Philipp Koehn and Josh Schroeder as part of Europarl v6 Preprocessing Tools suit[6]. The tool uses punctuation and capitalization clues to split paragraphs of sentences into files with one sentence per line. Fourthly, we performed tokenization using the same script as in Chapter 2, the script tokenizer.perl from Europarl v6 Preprocessing Tools suit. Finally, we corrected the automatic tools errors and removed the remaining noise manually. We left 1000 sentences in the new test set, with the total number of tokens 22498 so the size of this new test set is comparable with the UMC corpus test set (see the Table 3.5).

After that we converted the set to lower case and translated with MOSES decoder using the 4-gram models (trained with and without Wiki-dictionary) that were compared in the previous subsections.

[5] `http://code.google.com/p/gwtwiki/`

[6] `https://victorio.uit.no/langtech/trunk/tools/alignment-tools/europarl/`

| | UMC test set | Wiki test set |
|---|---|---|
| **numb. of sentences** | 1,000 | 1000 |
| **tokens** | 19,019 | 22,498 |
| **vocabulary size** | 7,001 | 7,694 |
| **average sentence length** | 19,02 | 22,50 |
| **standard dev. of sent. length** | 9,43 | 15,89 |

Table 3.5: Comparison of UMC and Wiki test sets' statistics

### 3.2.5 OOV analysis (Wiki-set)

The number of OOV words on the new test set was 2260 (3537 in total, if we count repeated OOV words) for the model that was trained just on the training set of the UMC corpus. For the model trained on the training set of the UMC corpus with Wiki-dictionary the number of OOV on the new test set was 1878 (2342 in total, if we count repeated OOV words). These numbers are twice bigger than the number of OOV in the previous experiment with UMC corpus test set while the sizes of the test sets were comparable, see Table 3.6 for the details. The increase of the number of OOV is mostlikely caused by the shift of the topic: most sentences from the UMC test set were related to politics and medical research while the main topics of the Wiki-set were biographies of popular singers and politician.

The diagrams on the Figure 3.6 show the distribution of OOV along the Wiki-set. We see that the diagrams have similar shape while the numbers of OOV words

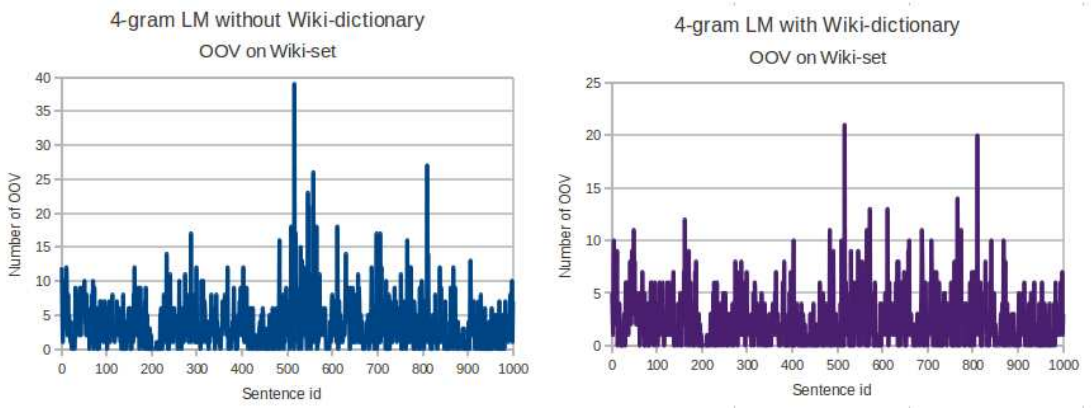



Figure 3.6: Distribution of OOV words on the Wiki-set for the 4-gram model trained without Wiki-dictionary (on the left) and for the 4-gram model trained with Wiki-dictionary (on the right)

for the model trained with Wiki-dictionary are much lower. The highest number of OOV words are between sentences number 500 and 830. This part of the Wiki-set is about the biography of Vladimir Lenin and his political activity. This text contains many terms that are specific for that particular historical period. The fact that this terminology is very specific explains the higher number of OOV words. The number of OOV words by the end of the Wiki-set is lower because the last part of the corpus talks about the desease and death of Vladimir Lenin,

social opinion about this person and representation of his character in arts and language (films, anecdotes). The language of this part of the corpus is closer to journalistic, and the last part of the article doesn't contain many special political and historical terms.

| | **UMC test set** | | **Wiki test set** | |
|---|---|---|---|---|
| | **4-gram model w/t Wiki-dict.** | **4-gram model with Wiki-dict** | **4-gram model w/t Wiki-dict.** | **4-gram model with Wiki-dict** |
| **# of tokens** | 19,019 | | 22,498 | |
| **# of OOV** | 988 | 945 | 3,537 | 2,342 |
| **vocabulary size** | 7,001 | | 7,694 | |
| **# of diff. OOV** | 934 | 906 | 2,260 | 1,878 |

Table 3.6: Comparison of OOV words statistics for the 4-gram models trained with and without Wiki-dictionary on UMC and Wiki test sets

Also some terminology was better translated with the model that was trained with Wiki-dictionary. For example, the translation of the 81st sentence of the Wiki-set given by the model trained without Wiki-dictionary contained 3 OOV words:

продюсер, мадонны, хит
*Transliteration: prodjuser, madonny, xit*

while with the second model didn't have any OOV words on the same sentence. One of this words (мадонны - madonna, transliteration: madonny) is a named entity while two others (продюсер - producer, transliteration: prodjuser; хит - hit, transliteration: hit) are just special terms that are among the most commonly used words in the popular musical industry.

### 3.2.6 Manual Ranking (Wiki-set)

We did manual evaluation of the translations that were produced by the 4-gram language model trained without Wiki-dictionary and by the 4-gram language model trained with Wiki-dictionary. Unfortunately, we couldn't use the same approach as in section to get a sample from the Wiki set because due to the high number of OOV words in each sentence it is too difficult to rank these sentences. Therefore, we wrote a script that chose such pairs of sentences that:

- both sentences had at most two OOV words (as there were only 92 sentences with at most two OOV words, we randomly chose the 8 sententeces from the set of such pairs that had exactly 3 OOV words in both sentences);
- at least one of the sentences was longer than 124 characters (counting spaces);
- the sentences were not equal.

We collected the last 8 pairs of sentences so that they have equal number of OOV words because we already did OOV analysis in the previous section and in this evaluation we are interested in all the aspects of the quality of the translation.

The condition about the lenght of the sentence was introduced because we are interested to compare longer sentences.

Although the chosen set of 100 sentences was not as uniformly distributed as the one from the section, but still it was more or less uniformly sampled, according to our observations.

We gave the sentence higher rank if it reflected the original meaning better,or had better grammatical structure, or less OOV words, or less missing words on other equal conditions. In about half of the cases OOV words were the main reason of ranking one system first against the other and the second most frequent reason for the final decision is correspondence of the the meaning of the translation to meaning of the original sentence. The results are summarized in the Table 3.7

| | **Model w/t Wiki-dict is ranked first** | **Model with Wiki-dict is ranked first** | **Translations are equally bad/good** |
|---|---|---|---|
| **ranking of first 50 sentences** | 23 | 27 | 0 |
| **ranking of 100 sentences** | 44 | 50 | 6 |

Table 3.7: The results of the manual ranking of the translations produced by the 4-gram language model trained without Wiki-dictionary and the 4-gram language model trained with Wiki-dictionary on the sample from the WIKI-set

One half of the sample the model trained with Wiki-dictionary was ranked first and for 6 sentences the models were ranked equally. After comparison of the first 50 sentences we got less convincing results which could mean that on the first half of the sample models behaved similarly. This evaluation gives some evidence that model trained with Wiki-dictionary translates better on the Wiki-set.

As we expected there were many cases when the model trained without Wiki-dictionary didn't translate named entities while the model trained with Wiki-dictionary recognized and translated the named entities correctly.

```
<s1>after the death of фредди меркьюри remaining members groups ,
using records his votes , could provoke in 1995 , the latest
krylenko queen - made in heaven .  </s1>
<s2>after the death of freddie mercury remaining members of
groups , through records his votes , managed to issue in 1995 , the
last queen made in heaven .  </s2>

<s1>sociologist at yale university immanuel валлерстайн believes
that by 2050 , lenin inevitably become a national hero russia .
</s1>
<s2>marketing sociology at yale university , immanuel wallerstein
believes that by 2050 , lenin inevitably will be the national hero
russia .  </s2>
```

```
<s1>спиридович а.  и.  history большевизма in russia :  from arise
when .  - paris , 1922 .  с.  - 137 .  </s1>
<s2>alexander spiridovich ivan alexander history bolshevism in
russia :  from arise , prior to power .  - paris , 1922 .  - s.  137
.
</s2>
```

### 3.2.7 Evaluation Without Comments In Brackets (UMC Test Set)

We performed another experiment to assess the impact of the comments in brackets for the dictionary phrases like

*Damo (philosopher) | Damo (Greece)*
**Дамо (философ)**
*Transliteration: damo (filosof)*

on the BLEU score and the number of the OOV words when we use the Wiki-dictionary for the machine translation. We filtered out these comments and incorporated the dictionary into the UMC training set.

The BLEU score for the new model was 20.89 while the BLEU score for the model trained on the training set with Wiki-dictionary with comments was 20.42 and the BLEU score for the model trained on the training set without Wiki-dictionary was 21.42 (see Table 3.1). The new score didn't beat the score of the model trained without the Wiki-dictionary and the difference between the models that utilized Wiki-dictionary for training is rather small: with the paired bootstrap re-sampling we got 87.1% statistical significance that the new model was better.

The total number of different OOV words for the new model was 889 (929 in total if we count repeated OOV words) which is a bit less than for the model that was trained with Wiki-dictionary with comments that had 906 different OOV words and 945 OOV words in total (see section 3.2.1).

These results provide some evidence that translation can be slightly improved by removing the comments in brackets from the dictionary. The comments are useful for word sense disambiguation and for named entity recognition and classification but they might introduce noise in the phrase table of machine translation system. The comments may cause the wrong alignment of the words. Probably, we could reach more significant improvement of the BLEU score if we could disambiguate the cases when the comma is a part of the main part of the English title and when the comma (instead of brackets) separates the main part of the title from the comment.

# 4. Conclusion

In this work we attempted to create and evaluate bilingual bidirectional English-Russian dictionaries derived from Wikipedia article titles. Due to the peculiarity of Wikipedia content these dictionaries differ from traditional dictionaries, e.g. most of the phrases and words are named entities (personal names, geographical names, titles of the movies etc.) and at least 96% of the basic terminology that Wiki-dictionaries share with Mueller's English-Russian dictionary are noun phrases.

The main goals of this study were to develop a dictionary, evaluate its quality and examine its content and apply it for the task of statistical machine translation.

We performed named entity recognition and classification in order to learn more about the content of the dictionary. Having this knowledge we define for which domains this dictionary could be most suitable. Our sub-goal was to evaluate if the heuristics for NER and algorithm for NEC provide results of sufficient quality.

Due to the nature of the Wiki-dictionaries, they contain specific terminology and named entities that cannot be found in the traditional dictionaries. However, the recall of the basic vocabulary computed against the machine-readable version of the Mueller's English-Russian dictionary was much lower our intuitive estimations: the highest value we got was only 7.42%. Moreover, even the recall of the names, geographical names and especially for abbreviations did not correspond to our expectations.

The evaluation was also performed on the parallel corpus which was created at the Charles University in Prague. We found out that translation pairs from the Wiki-dictionary occur in the corpus, however, there is a significant number of sentences that don't contain any terms from Wiki-dictionary.

As the next step, we incorporated the Wiki-dictionary into the training set of the corpus and trained language model. We compared the results with the translations that we got with the language model trained without Wiki-dictionary. According to the BLEU score, paired bootstrapping, OOV words analysis and manual evaluation, the accuracy dropped down. The difference in the domain of the corpus and Wiki-dictionary could explain such result. We collected a new test set from the texts of three Wikipedia articles that we called Wiki-set and compared the models on this set. According to OOV words analysis and manual evaluation, the model trained with Wiki-dictionary performed better on the Wiki-set. The high number of OOV words that we got in our experiments with machine translation system could be explained by the fact, that Russian language has rich inflection and the words that are present in the training set and in the dictionary, might have different forms in the test set and therefore cannot be recognized.

As a future work, the dictionary should be evaluated on the parallel corpus from another domain such as newspaper articles about show-business news or popular science as we consider that the dictionary is appropriate for such kind of texts. We would have to run the machine translation experiments on the such corpus to see if we can get any improvements of the translation quality.

We would like to investigate further if the comments in brackets that occur in some titles have any statistically significant influence on the output of machine

translation system trained on the new domain.

Russian language has rich morphology therefore we would like to connect the dictionary to the morphological analyzer to obtain and predict all possible forms of the words in the dictionary. This information might be also useful for the experiments with factored machine translation.

We also consider the improvement of the named entity recognition, we would like to use more sophisticated techniques to distinguish terminology from named entities.

# Bibliography


[1] Bertoldi, N.; Haddow, B.; Fouet, J.-B.: Improved Minimum Error Rate Training in Moses. *The Prague Bulletin of Mathematical Linguistics*, 2009: s. 7–16.

[2] Brown, P. F.; Cocke, J.; Pietra, S. A. D.; aj.: A statistical approach to machine translation. *Comput. Linguist.*, June 1990: s. 79–85, ISSN 0891-2017.

[3] Brown, P. F.; Della Pietra, S. A.; Della Pietra, V. J.; aj.: But dictionaries are data too. In *Proceedings of the workshop on Human Language Technology*, HLT '93, Stroudsburg, PA, USA: Association for Computational Linguistics, 1993, ISBN 1-55860-324-7, s. 202–205.

[4] Brown, P. F.; Pietra, V. J.; Pietra, S. A. D.; aj.: The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 1993: s. 263–311.

[5] Bunescu, R. C.; Pasca, M.: Using Encyclopedic Knowledge for Named entity Disambiguation. In *EACL*, 2006.

[6] Callison-Burch, C.; Fordyce, C.; Koehn, P.; aj.: (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic: Association for Computational Linguistics, June 2007, s. 136–158.

[7] Dakka, W.; Cucerzan, S.: Augmenting Wikipedia with Named Entity Tags. In *Proceedings of IJCNLP 2008*, 2008.

[8] van der Eijk, P.; Bloksma, L.; van der Kraan, M.: Towards developing reusable NLP dictionaries. In *Proceedings of the 14th conference on Computational linguistics - Volume 1*, COLING '92, Stroudsburg, PA, USA: Association for Computational Linguistics, 1992, s. 53–59.

[9] Erdmann, M.; Nakayama, K.; Hara, T.; aj.: An approach for extracting bilingual terminology from Wikipedia. In *Proceedings of the 13th international conference on Database systems for advanced applications*, DASFAA'08, Berlin, Heidelberg: Springer-Verlag, 2008, ISBN 3-540-78567-1, 978-3-540-78567-5, s. 380–392.

[10] Erdmann, M.; Nakayama, K.; Hara, T.; aj.: A Bilingual Dictionary Extracted from the Wikipedia Link Structure. In *DASFAA'08*, 2008, s. 686–689.

[11] Gelb, I.: Lexicography, lexicology, and the Accadian dictionary. In *A. Andre Martinet e structuralismo e Historia II*, Canarias: Universidad de la Laguna, 1958, str. 66.

[12] Habash, N.; Saleh, I.: Automatic extraction of lemma-based bilingual dictionaries for morphologically rich languages. 2009, in the Third Workshop on Computational Approaches to Arabic Script-based Languages at the Machine Translation Summit XII Ottawa, Ontario, Canada.

[13] Hildebrand; Eck, M.; Vogel, S.; aj.: Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval. In *Proceedings of EAMT*, Budapest, Hungary, May 2005.

[14] Héja, E.: Automatic Methods in the Construction of Bilingual Dictionaries. 2008.

[15] Kazama, J.; Torisawa, K.: Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In *EMNLP-CoNLL'07*, 2007, s. 698–707.

[16] Klyueva, N.; Bojar, O.: UMC 0.1: Czech-Russian-English Multilingual Corpus. In *Proc. of International Conference Corpus Linguistics*, October 2008, s. 188–195.

[17] Knopp, J.: Classification of named entities in a large multilingual resource using the Wikipedia category system. *Master Thesis*, 2010.

[18] Koehn, P.: Statistical Significance Tests for Machine Translation Evaluation. In *EMNLP*, 2004, s. 388–395.

[19] Koehn, P.; Hoang, H.; Birch, A.; aj.: Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL*, The Association for Computer Linguistics, 2007.

[20] Koehn, P.; Och, F. J.; Marcu, D.: Statistical phrase-based translation. In *NAACL '03*, Morristown, NJ, USA: Association for Computational Linguistics, 2003, s. 48–54.

[21] Kolovratník, D.; Klyueva, N.; Bojar, O.: Statistical Machine Translation between Related and Unrelated Languages. In *ITAT 2009 Information Technologies – Applications and Theory*, September 2009.

[22] Longman: *Longman Dictionary of Contemporary English*. Burnt Mill, Harlow, Essex, England: Longman House, 1987.

[23] Manning, C. D.; Schütze, H.: *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999, ISBN 0262133601.

[24] de Melo, G.; Weikum, G.: Untangling the Cross-Lingual Link Structure of Wikipedia. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010) : proceedings of the conference*, Uppsala, Sweden: Association for Computational Linguistics, 2010, ISBN 978-1-932432-67-1, s. 844–853.

[25] Munteanu, D. S.; Marcu, D.: Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 2005: s. 477–504.

[26] Nygaard, L.; Lønning, J. T.; Oepen, S.: Using a bi-lingual dictionary in lexical transfer. In *In Proceedings of the 11th conference of the European Asoociation of Machine Translation*, 2006.

[27] Och, F. J.; Ney, H.: A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, Stroudsburg, PA, USA: Association for Computational Linguistics, 2000, s. 1086–1090.

[28] Och, F. J.; Ney, H.: A systematic comparison of various statistical alignment models. *Comput. Linguist.*, March 2003: s. 19–51, ISSN 0891-2017.

[29] Richman, A. E.; Schone, P.: Mining Wiki Resources for Multilingual Named Entity Recognition. In *Proceedings of ACL-08: HLT*, Columbus, Ohio: Association for Computational Linguistics, June 2008, s. 1–9.

[30] Rohit Bharadwaj G, V. V., Niket Tandon: An Iterative approach to extract dictionaries from wikipedia for under-resourced languages. 2010, iCON 2010, IIT Kharagpur, India.

[31] Sadat, F.; Yoshikawa, M.; Uemura, S.: Learning Bilingual Translations from Comparable Corpora to Cross-Language Information Retrieval: Hybrid Statistics-based and Linguistics-based Approach. In *In Proc. IRAL 2003*, 2003.

[32] Stanley F. Chen, J. G.: An Empirical Study of Smoothing Techniques for Language Modeling. In *Proc. of ACL*, 1996, s. 310–318.

[33] Tjong Kim Sang, E. F.; De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*, editace W. Daelemans; M. Osborne, Edmonton, Canada, 2003, s. 142–147.

[34] Wehrli, E.; Nerima, L.; Scherrer, Y.: Deep Linguistic Multilingual Translation and Bilingual Dictionaries. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece: Association for Computational Linguistics, March 2009, s. 90–94.

# List of Tables


2.1 Confusion matrix . . . . . 15
2.2 Accuracy, Precision, Recall, F1-score . . . . . 15
2.3 NER of the Wiki-dictionary entries . . . . . 19
2.4 Evaluation of NER on the sample from Wiki-dictionary . . . . . 19
2.5 Named entity types . . . . . 21
2.6 NEC of the named entities from the Wiki-dictionary . . . . . 22
2.7 Changes in NEC of the Wiki-dictionary after corrections . . . . . 24
2.8 Evaluation of the NER for Wiki-dictionary . . . . . 25
2.9 Evaluation of the NERC for Wiki-dictionary . . . . . 25
2.10 Evaluation of the named entity label PER for Wiki-dictionary . . 25
2.11 Evaluation of the named entity label LOC for Wiki-dictionary . . 26
2.12 Evaluation of the named entity label ORG for Wiki-dictionary . . 26
2.13 Evaluation of the named entity label MISC for Wiki-dictionary . . 27
2.14 Evaluation of the NERC for English-German Wiki-dictionary . . . 27
2.15 Size of Mueller's Dictionaries . . . . . 29
2.16 Statistics of Wiki-dictionary with redirects on Mueller Dictionaries 32
2.17 Statistics of noun phrases from Wiki-dict. in Mueller's base dict. . 33
2.18 Statistics of English part of Wiki-dictionary on Mueller Geogr. . . 34
2.19 Statistics of Wiki-dictionary w/t synonyms on Mueller Dictionaries 34
2.20 UMC corpus statistics . . . . . 36
2.21 Pair statistics of Wiki-dictionaries on the UMC train set . . . . . 37
2.22 Pair statistics of Wiki-dictionaries on the UMC development set . 38
2.23 Pair statistics of Wiki-dictionaries on the UMC test set . . . . . 39

3.1 BLEU score for the trained models . . . . . 47
3.2 Individual and cumulative n-gram scoring . . . . . 47
3.3 Paired boostrap re-sampling on the samples of the UMC test set . 50
3.4 Manual ranking on the sample from the UMC test set . . . . . 51
3.5 Comparison of UMC and Wiki test sets' statistics . . . . . 52
3.6 OOV words statistics on the UMC test set and on the Wiki-set . . 53
3.7 Manual ranking on the sample from the WIKI test set . . . . . 54

C.1 Pair statistics without comments in brakets on the UMC train set 81
C.2 Pair statistics without comments in brakets on UMC dev. set . . 82
C.3 Pair statistics without comments in brakets on the UMC test set . 83

# List of Figures

1.1 The architecture of MT system . . . . . 3

2.1 Sample key category phrases . . . . . 16
2.2 Results for the NERC task . . . . . 17
2.3 Words that were considered non-content in the NER experiment . 18
2.4 Comments used to correct the results of NERC . . . . . 23
2.5 Recall of the Wiki-dict. on Mueller dict. . . . . 32
2.6 Recall of the Wiki-dict. without synonyms on Mueller dict. . . . . 35

3.1 Noisy channel . . . . . 40
3.2 Language model and translation model . . . . . 40
3.3 Decoder . . . . . 41
3.4 Basic components of Moses toolkit . . . . . 45
3.5 Distribution of OOV words on the UMC test set . . . . . 49
3.6 Distribution of the OOV words on the Wiki-set . . . . . 52

# List of Abbreviations

**I (EN-DE-RU)** Wiki-dictionary created from the intersection of the interlanguage links between English, Russian and German languages

**I+R (EN-DE-RU)** Wiki-dictionary created from the intersection of both interlanguage links and redirect pages between English-Russian, English-German and German-Russian dictionaries of the type I+R

**I+R+T** Wiki-dictionary created from the the dictionary of the type I+R plus the new English-Russian links obtained by transitivity from English-German and German-Russian dictionaries

**I+R** Wiki-dictionary created using interlanguage links and redirect pages

**I** Wiki-dictionary created using only interlanguage links

**LOC** named entity type “location”

**MISC** named entity type “miscellaneous”

**NEC** named entity classification

**NERC** named entity recognition and classification

**NER** named entity recognition

**NE** named entity

**OOV** out-of-vocabulary

**ORG** named entity type “organization”

**PER** named entity type “person”

**UMC** ÚFAL Multilingual Corpus

# A. Sample from English-Russian Wiki-dictionary

*1832 in rail transport*
**1832 год в истории железнодорожного транспорта**

*Abja Parish*
**Абья (волость)|Абья**
LOC

*Aditya Mehta/Aditya Snehal Mehta*
**Мехта, Адитъя|Адита Мехта|Мехта, Адита|Адитъя Мехта**
PER

*Alejandro Toledo/Alejandro Toledo Manrique*
**Толедо, Алехандро|Алехандро Толедо**
PER

*Alex de Angelis/Alex De Angelis/Alex de angelis*
**Де Анджелис, Алекс|Алекс де Анджелис**
PER

*Alfa Laval*
**Alfa Laval|Альфа Лаваль**
ORG

*Alfara de la Baronia/Alfara de algimia/Alfara de Algimia*
**Альфара-де-Альхимиа**
LOC

*Alluvioni Cambiò/Alluvioni Cambio*
**Аллувьони-Камбио**
LOC

*Ambleside/Ambleside Church of England Primary School*
**Эмблсайд**
LOC

*Andreas Wolf*
**Вольф, Андреас|Андреас Вольф|Андреас Волф|Волф, Андреас|Вольф Андреас**
PER

*André Martinet/Andre Martinet*
**Мартине, Андре|Андре Мартине|Мартине А.|Мартине Андре**
PER

*Andy García/Andy Garcia*
**Гарсиа, Энди|Энди Гарсия|Энди Гарсиа|Гарсиа Э.|Гарсиа Энди**
PER

*Angel (season 4)*
**Ангел (сезон 4)**
MISC

*Anna Karenina (1967 film)*
**Анна Каренина (фильм, 1967)**
MISC

*Asahi, Mie*
**Асахи (Миэ)**
LOC

*Atul Dodiya/Atul dodiya*
**Додиа, Атул|Атул Додиа**
PER

*Bad Kohlgrub*
**Бад-Кольгруб**
LOC

*Banda people/Banda (CAR)/Banda (Sudan)*
**Банда (народ)**

*Bargarh district/Baragarh district/Baragarh District/Bargarh District*
**Баргарх (округ)**
LOC

*Beak/Beaks/Rhamphotheca/Bird's beak*
**Клюв|Rostrum**

*Beaulieu, Indre*
**Болье (Эндр)**
LOC

*Bedford, Wyoming/Bedford, WY*
**Бедфорд (Вайоминг)**
LOC

*Beerwah, Queensland*
**Бирва**
LOC

*Bieruń/Bierun*
**Берунь**
LOC

*Birch and Swinnerton-Dyer conjecture*
**Гипотеза Берча и Свиннертона-Дайера**

*Blizzard Beasts*
**Blizzard Beasts**
MISC

*Bonate Sopra*
**Бонате-Сопра**
LOC

*Boracite/Stassfurtite*
**Борацит**

*Braxton Bragg*
**Брэгг, Брэкстон|Брэкстон Брэгг**
PER

*Brazil at the 1964 Summer Olympics*
**Бразилия на летних Олимпийских играх 1964**

*Breakstep*
**Брейкстеп**

*Bruschetta/Bruchetta/Fettunta/Brushetta*
**Брускетта**

*Calw (district)/Calw district/Landkreis Calw*
**Кальв (район)**
LOC

*Charles Dupuy/Charles-Alexandre Dupuy/Charles Alexandre Dupuy*
**Дюпюи, Шарль|Шарль Дюпюи|Дюпюи Ш.|Дюпюи Шарль**
PER

*Chobits/Persocom/Chii-san/CHOBITS/Chobitsu/Persocon/Chobittsu*
**Чобиты|Chobits|Персокон|Persocon|Персоком|Chobit**
MISC

*Christian Pander*
**Пандер, Кристиан|Кристиан Пандер**
PER

*Châtenay-en-France|Chatenay-en-France*
**Шатене-ан-Франс**
LOC

*Conflenti*
**Конфленти**
LOC

*Copsi|Copsig|Copsi of Northumbria*
**Копси**
PER

*Crimolois*
**Кримолуа**
LOC

*Curiglia con Monteviasco*
**Курилья-кон-Монтевьяско**
LOC

*Damo (philosopher)|Damo (Greece)*
**Дамо (философ)**
PER

*Dangtu County|Dangtu*
**Данту**
LOC

*Dois Riachos*
**Дойс-Риашус**
LOC

*Eclipse (1962 film)|L'Eclisse|L' Eclisse|L'eclisse*
**Затмение (фильм, 1962)|L'Eclisse**
MISC

*Emanuel Berg*
**Берг, Эмануэль|Эмануэль Берг**
PER

*EncFS|Encfs*
**EncFS|Encfs**
MISC

*Erwin Olaf*
**Эрвин Олаф|Ирвин Олаф**
PER

*European Pairs Speedway Championship*
**Чемпионат Европы по спидвею среди пар**
MISC

*Fatima/Fatina/Fatima Az-Zehra*
**Фатима|Фатима (имя)**
PER

*Ford C170 platform/Ford CW170 platform/C170*
**Ford C170**
MISC

*Georg Jauer*
**Яуер, Георг|Георг Яуер**
PER

*Gmina Krośnice/Krośnice Commune/Gmina Krosnice/Krosnice Commune*
**Кроснице (гмина)|Гмина Кроснице|Кроснице**
LOC

*Gmina Mieścisko/Mieścisko Commune/Gmina Miescisko/Miescisko Commune*
**Месциско (гмина)|Гмина Месциско**
LOC

*Gmina Nowosolna/Nowosolna Commune*
**Новосольна (гмина)|Гмина Новосольна**
LOC

*Gmina Witnica/Witnica Commune*
**Витница (гмина)|Гмина Витница**
LOC

*Gmina Ziębice/Ziębice Commune/Gmina Ziebice/Ziebice Commune*
**Зембице (гмина)|Гмина Зембице**
LOC

*Gottfried Benn/Gottfried Ben*
**Бенн, Готфрид|Готфрид Бенн**
PER

*Gudea/Gudea of Lagash/Gudea cylinders*
**Гудеа**
PER

*Gyaros/Gyarus*
**Ярос**
LOC

*Hadoop/Apache Hadoop*
**Hadoop|Apache Hadoop**
MISC

*Handbuch des Schachspiels*
**Handbuch des Schachspiels|Хандбух**
MISC

*Henryk Hryniewski/Henryk Hrynievski*
**Гриневский, Генрик|Гриневский, Генрих|Гриневски, Генрих**
PER

*Hetzerath/Hetzerath (Eifel)*
**Хетцерат (Айфель)|Хетцерат**
LOC

*Ida Aalberg/Aalberg, Ida*
**Аалберг, Ида**
PER

*Infinity (Journey album)*
**Infinity (альбом)**
MISC

*Innerwish*
**InnerWish**
MISC

*Iraí de Minas*
**Ираи-ди-Минас**
LOC

*James Jones (basketball player)/James Andrew Jones*
**Джонс, Джеймс (баскетболист)**
PER

*Jayawijaya Regency*
**Джаявиджая|Джайавиджая**
LOC

*Joe Spence (footballer born 1898)*
**Спенс, Джо|Джо Спенс|Спенс Джо**
PER

*John Burr Williams*
**Уильямс, Джон Бэрр|Джон Бэрр Уильямс**
PER

*John Deacon/Deaky/John Richard Deacon*
**Дикон, Джон|Джон дикон|Джон Дикон|Дикон Д.|Дикон Джон**
PER

*Jozef Krnáč/Jozef Krnac*
**Крнач, Йозеф|Йозеф Крнач**
PER

*Kelsterbach*
**Кельстербах|Кельштербах**
LOC

*Kemerovo*
**Кемерово|Щегловск**
LOC

*Kemialliset Ystävät/Kemialliset Ystaevaet/Kemialliset Ystavat*
**Kemialliset Ystävät**
MISC

*Khaleeji (currency)/Gulf currency/Gulfo*
**Халиджи (валюта)|Калиджи (валюта)**

*Lequio Tanaro*
**Лекуйо-Танаро**
LOC

*Leo Jogiches/Jogiches*
**Йогихес, Лео|Лео Йогихес|Ян Тышка|Тышка, Ян|Йогихес|Йогихес Лео**
PER

*Mali at the 1968 Summer Olympics*
**Мали на летних Олимпийских играх 1968**

*Marigold (2007 film)*
**Мариголд: Путешествие в Индию (фильм)**
MISC

*Mazzarrà Sant'Andrea/Mazzarra Sant'Andrea*
**Маццарра-Сант'Андреа**
LOC

*Merida/Mérida (city)/Merida (city)/Mèrida*
**Мерида**
LOC

*Mertendorf/Mertendorf (bei Naumburg)*
**Мертендорф (Наумбург)**
LOC

*Michael Asen II of Bulgaria/Michael Asen II*
**Михаил II Асень**
PER

*Miguel Montuori*
**Монтуори, Мигель|Мигель Монтуори|Монтуори|Монтуори Мигель**
PER

*Miltzow (Amt)/Amt Miltzow*
**Управление Мильцов**
LOC

*Mount Hillaby*
**Хиллаби**
LOC

*Myrmothera*
**Мирмотеры|Myrmothera**

*Nomi (TN)/Nomi, Italy*
**Номи (провинция Тренто)**
LOC

*Novo Konomladi*
**Ново-Кономлади**
LOC

*OpenFOAM/FreeFOAM*
**OpenFOAM**
MISC

*Operation Dumbo Drop/Dumbo drop*
**Операция «Дамбо» (фильм)|Операция «Дамбо»**
MISC

*Paderno Franciacorta*
**Падерно-Франчакорта**
LOC

*Paul Le Guen*
**Ле Гуэн, Поль|Поль Ле Гуэн**
PER

*Pavel Mareš/Pavel Mares*
**Мареш, Павел|Павел Мареш|Мареш Павел|Мареш П.**
PER

*Pedra Bela*
**Педра-Бела**
LOC

*Polesia Voblast/Voblasts of Polesia/Polesie Voblast*
**Полесская область|Полесская Область**
LOC

*Prone Lugo AD/Azkar Lugo FS*
**Аскар Луго|Азкар Луго**
ORG

*Roman Butenko*
**Бутенко, Роман Анатольевич**
PER

*Ruslan Alekhno/Ruslan Alehno*
**Алехно, Руслан Фёдорович|Руслан Алехно|Алехно, Руслан**
PER

*Sabunçu, Baku/Sabuncu, Baku*
**Сабунчи**
LOC

*Saints Peter and Paul Cathedral (Kazan)*
**Петропавловский собор (Казань)**
LOC

*Sam Houser/Sam houser*
**Хаузер, Сэм|Сэм Хаузер**
PER

*Samantha (film)/Samantha (movie)*
**Саманта (фильм)**
MISC

*Saxon (band)*
**Saxon**
MISC

*Senhora de Oliveira/Senhora de oliveira*
**Сеньора-ди-Оливейра**
LOC

*Setuid/Setgid*
**Suid**

*Smuggling/Smuggle/Smugglers/Trafficked/Smuggler/Smuggled*
**Контрабанда|Контрабандист**

*Stecklenberg*
**Штекленберг**
LOC

*Stefan Toshev*
**Тошев, Стефан|Стефан Тошев|Тошев Стефан**
PER

*Surplice/Surplices/Superpelliceum/Surpliced/Surplicing*
**Комжа|Стихарь католический|Стихарь (католицизм)**

*Tefé River*
**Тефе (река)**
LOC

*Tupolev Tu-85*
**Ту-85**
MISC

*Turkey at the 1976 Summer Olympics*
**Турция на летних Олимпийских играх 1976**

*Ugo Locatelli*
**Локателли, Уго|Уго Локателли**
PER

*Uharte-Arakil*
**Уарте-Аракиль**
LOC

*Urucuia*
**Урукуя**
LOC

*Vandenesse-en-Auxois*
**Ванденес-ан-Осуа**
LOC

*Vekhi*
**Вехи (сборник)|Вехи. Сборник статей о русской интеллигенции**
MISC

*Walther Funk/Walter Funk/Walther Emmanuel Funk*
**Функ, Вальтер|Вальтер Эмануэль Функ|Вальтер Функ|Функ Вальтер**
PER

*Wernau/Wernau (Neckar)*
**Вернау (Неккар)|Вернау**
LOC

*Yutaka Higuchi*
**Хигути, Ютака|Ютака Хигути|Хигути Ютака**
PER

# B. Sample from English-German Wiki-dictionary

*1930 World Ice Hockey Championships | 1930 Men's World Ice Hockey Championships*
**Eishockey-Weltmeisterschaft 1930**
MISC

*3rd arrondissement of Paris | 3eme arrondissement, Paris | 3ème arrondissement, Paris | 75003 | IIIe arrondissement | IIIe arrondissement (Paris) | Paris 3e | The 3rd arrondissement of Paris*
**3. Arrondissement (Paris) | 3. Pariser Arrondissement | Arrondissement du Temple**
LOC

*9 | 762 AUC | 9 (year) | 9 A.D. | 9 AD | 9 CE | AD 9*
**9**
MISC

*Adalbert I, Margrave of Tuscany | Adalbert, Duke of Tuscany | Adalbert I of Tuscany*
**Adalbert I. (Tuszien)**
PER

*Aras River | Arak's | Araks | Araks River | Araks river | Aras river | Arax | Arax River | Arax river | Araxes | Araxes River | Araxes river | Araz | Araz River | Araz river*
**Aras (Fluss)**
LOC

*Beak | Beaks | Bird's beak | Rhamphotheca*
**Schnabel | Oberschnabel | Unterschnabel**

*Bernd Schneider (racing driver) | Bernd Schneider (auto racing) | Bernd Schneider (racecar driver)*
**Bernd Schneider (Rennfahrer)**
PER

*Calw (district) | Calw district | Landkreis Calw*
**Landkreis Calw | Kreis Calw**
LOC

*Christian Pander*
**Christian Pander | Funky Pee**
PER

*Citrulline | C6H13N3O3 | Citrulline malate*
**Citrullin | L-Citrullin**

*Claude Auchinleck | Auchinleck, Sir Claude | Claude E. Auchinleck | Claude J. E. Auchinleck | Claude John Eyre Auchinleck | Field Marshal Auchinleck | Field Marshall Auchinleck | General Auchinleck | General Auckinleck | Sir Claude Auchinleck | Sir Claude John Eyre Auchinleck*
**Claude Auchinleck**
PER

*Commandry (feudalism) | Commandary | Commandery (feudalism) | Commandries | Commandry*
**Kommende | Commanderie | Comturei | In commendam | Kommendatar | Kommendatarabt | Komturei | Komturhof | Laienabt**

*Dürrenroth | Duerrenroth | Duerrenroth, Switzerland | Duerrenroth (Bern) | Duerrenroth (Berne) | Duerrenroth BE | Durrenroth | Durrenroth, Switzerland | Durrenroth (Bern) | Durrenroth (Berne) | Durrenroth BE | Dürrenroth, Switzerland | Dürrenroth (Bern) | Dürrenroth (Berne) | Dürrenroth BE*
**Dürrenroth**
LOC

*Eccentricity (mathematics) | Eccentricity (math) | First eccentricity | Linear eccentricity | Second eccentricity | Third eccentricity*
**Exzentrizität (Mathematik) | Lineare Exzentrizität | Numerische Exzentrizität**

*Eclipse (1962 film) | L'Eclisse | L' Eclisse | L'eclisse*
**Liebe 1962**
MISC

*F. Sionil José | F.Sionil Jose | F. Sionil Jose | F sionil jose | Francisco Sionil Jose | Francisco Sionil José*
**F. Sionil José | F. Sionil Jose | Francisco Sionil Jose | Francisco Sionil José**
PER

*Gudea | Gudea cylinders | Gudea of Lagash*
**Gudea | Gudea von Lagasch | Gudeä**
PER

*Gyaros | Gyarus*
**Gyaros | Giaros | Gyáros**
LOC

*Heinrich Wilhelm Matthäus Olbers | H. Olbers | H. W. M. Olbers | H. W. Olbers | Heinrich Olbers | Heinrich W. M. Olbers | Heinrich W. Olbers | Heinrich Wilhelm Matthaeus Olbers | Heinrich Wilhelm Matthaus Olbers | Heinrich Wilhelm Matthias Olbers | Heinrich Wilhelm Olbers | Olbers | Wilhelm Olbers*
**Heinrich Wilhelm Olbers | Heinrich Olbers | Wilhelm Olbers**
PER

*Holography | Hologram | Hologrammatic | Hologramme | Holograms | Holographic art | Holographic optics | Holographics*
**Holografie | Holo | Hologramm | Holographie | Holos | Regenbogenholographie**

*John II of Jerusalem | Jean I de Chypre | John I of Cyprus*
**Johann I. (Zypern)**
PER

*Kemerovo*
**Kemerowo | Kemerovo**
LOC

*Lattice (order) | Bounded lattice | Join (lattice theory) | Lattice (algebra) | Lattice (order theory) | Lattice order | Lattice theory | Meet (lattice theory)*
**Verband (Mathematik) | Distributiver Verband | Halbverband | Verbandstheorie | Vollständiger Verband**

*Lichtenau, Baden-Württemberg | Lichtenau, Baden | Lichtenau, Baden-Wuerttemberg | Lichtenau, Baden-Wurttemberg | Lichtenau, Rastatt | Lichtenau (Baden) | Lichtenau (Rastatt) | Lichtenau (Rastatt County)*
**Lichtenau (Baden) | Grauelsbaum | Muckenschopf | Scherzheim | Ulm (Lichtenau)**
LOC

*Line-crossing ceremony | Ancient Order of the Deep | Baptism on the line | Equatorial baptism | Golden shellback | Line crossing ceremony | Order of the Polar Bear | The Ancient Order of the Deep*
**Äquatortaufe | Linientaufe**
MISC

*Malnutrition*
**Mangelernährung**

*Medicinal mushrooms*
**Mykotherapie**

*Meteorology | High Pressure Systems | Hurricane specialist | List of weather-related disasters | Low Pressure Systems | Meteorological | Meteorologie | Meteorología | Metereologia | Metereology | Meterological | Meterologist | Meterology | Nuclear meteorology | Slight Risk Of Severe Thunderstorms*
**Meteorologie | Atmosphärenphysik | Meteorologe | Meteorologisch | Physik der Atmosphären | Strahlungsforschung | Strahlungsmessung | Wetterkunde | Wetterumschwung**

*Nassau, Germany | Nassau, Rhineland-Palatinate | Nassau (Lahn) | Nassau an der Lahn*
**Nassau (Lahn) | Nassau an der Lahn**
LOC

*Olympic Flame | Olympic Fire | Olympic Torch | Olympic cauldron | Olympic fire | Olympic flame | Olympic torch*
**Olympischer Fackellauf | Fackellauf | Olympische Fackel | Olympische Flamme**
MISC

*Schömberg, Calw | Schoemberg, Calw | Schoemberg (Landkreis Calw) | Schoemberg im Schwarzwald | Schomberg, Calw | Schomberg (Landkreis Calw) | Schomberg im Schwarzwald | Schömberg (Landkreis Calw) | Schömberg im Schwarzwald*
**Schömberg (Landkreis Calw) | Bieselsberg | Charlottenhöhe | Langenbrand (Schömberg) | Oberlengenhardt | Schwarzenberg (Schömberg)**
LOC

*Seoul Metropolitan Subway | Seoul Subway | Seoul subway | Seoul subway system*
**U-Bahn Seoul**
LOC

*Setuid | Setgid*
**Setuid | SUID | SetUID | SetUID-Bit**

*Smuggling | Smuggle | Smuggled | Smuggler | Smugglers | Trafficked*
**Schmuggel | Schmuggeln | Schmuggler**

*Tupolev Tu-85*
**Tupolew Tu-80/85 | Tu-80 | Tu-85 | Tupolew Tu-85**
MISC

*Valuation (algebra) | Maximal ideal of a valuation | Prime ideal of a valuation | Residue field of a valuation | Trivial valuation | Valuation group | Valuation ring of a valuation | Valuation theory | Value group | Valued field*
**Bewertungstheorie | Für p ganze Zahl | Für p gebrochene Zahl | P-Exponent**

*Vekhi*
**Wechi**
MISC

*Vincent de Paul | Church of St Vincent de Paul | Saint Vincent DePaul | Saint Vincent De Paul | Saint Vincent de Paul | Saint vincent de paul | St-Vincent-de-Paul | St. Vincent DePaul | St. Vincent De Paul | St. Vincent de Paul | St. Vincent of Paul | St Vincent de Paul | Vincent DePaul | Vincent De Paul | Vincent de Paul (saint) | Vincentius a Paolo*
**Vinzenz von Paul | St. Vincent de Paul**
PER

*Witch of Agnesi | Agnesi's Witch | Agnesi's witch | Agnesi Witch | Agnesi witch | Agnesienne | Agnésienne | Cubique d'Agnesi | Curve of Agnesi | Versiera | Versiera di Agnesi | Witch of Agnasi | Witch of Maria Agnesi | Witch of agnesi*
**Versiera der Agnesi | Agnesi-Kurve | Versiera der Maria Agnesi | Versiera di Agnesi**
MISC

# C. Corpus statistics

| | NUMBER OF SENTENCES | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (COUNTS BY WORD FORMS) | | | | | (COUNTS BY NORMAL FORMS) | | | | |
| # of pairs | I (EN-DE-RU) | I +R (EN-DE-RU) | I | I +R | I +R +T | I (EN-DE-RU) | I +R (EN-DE-RU) | I | I +R | I +R +T |
| 55 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 52 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 50 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 49 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 48 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 47 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 46 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 42 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 39 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 38 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 2 |
| 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 3 |
| 36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 4 |
| 35 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 3 | 2 |
| 34 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 2 |
| 33 | 0 | 0 | 0 | 0 | 0 | 5 | 4 | 1 | 7 | 8 |
| 32 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 7 | 8 | 6 |
| 31 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 6 | 7 | 7 |
| 30 | 0 | 0 | 1 | 1 | 1 | 1 | 2 | 9 | 2 | 5 |
| 29 | 0 | 0 | 1 | 3 | 3 | 3 | 2 | 7 | 12 | 15 |
| 28 | 0 | 0 | 1 | 1 | 3 | 1 | 1 | 11 | 12 | 12 |
| 27 | 2 | 2 | 3 | 3 | 1 | 2 | 2 | 11 | 16 | 17 |
| 26 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 20 | 24 | 28 |
| 25 | 1 | 1 | 1 | 1 | 1 | 13 | 13 | 24 | 28 | 28 |
| 24 | 1 | 1 | 4 | 6 | 7 | 6 | 6 | 38 | 43 | 62 |
| 23 | 1 | 1 | 8 | 7 | 6 | 14 | 14 | 58 | 69 | 63 |
| 22 | 2 | 2 | 6 | 5 | 6 | 9 | 9 | 60 | 74 | 87 |
| 21 | 5 | 5 | 2 | 3 | 3 | 21 | 21 | 97 | 120 | 129 |
| 20 | 4 | 4 | 12 | 14 | 15 | 34 | 34 | 102 | 139 | 148 |
| 19 | 10 | 10 | 12 | 14 | 14 | 57 | 59 | 178 | 192 | 206 |
| 18 | 8 | 8 | 14 | 18 | 23 | 65 | 68 | 217 | 273 | 285 |
| 17 | 10 | 10 | 32 | 43 | 44 | 110 | 109 | 324 | 373 | 389 |
| 16 | 16 | 16 | 45 | 48 | 55 | 160 | 163 | 379 | 440 | 490 |
| 15 | 20 | 22 | 73 | 101 | 108 | 191 | 191 | 521 | 585 | 643 |
| 14 | 48 | 48 | 130 | 156 | 166 | 285 | 287 | 735 | 861 | 881 |
| 13 | 65 | 66 | 202 | 206 | 214 | 514 | 518 | 982 | 1063 | 1082 |
| 12 | 94 | 94 | 259 | 305 | 322 | 659 | 655 | 1326 | 1426 | 1492 |

| | NUMBER OF SENTENCES | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (COUNTS BY WORD FORMS) | | | | | (COUNTS BY NORMAL FORMS) | | | | |
| # of pairs | I (EN-DE-RU) | I +R (EN-DE-RU) | I | I +R | I +R +T | I (EN-DE-RU) | I +R (EN-DE-RU) | I | I +R | I +R +T |
| 11 | 168 | 170 | 388 | 471 | 502 | 928 | 943 | 1693 | 1903 | 1998 |
| 10 | 293 | 291 | 620 | 703 | 739 | 1369 | 1376 | 2326 | 2605 | 2725 |
| 9 | 420 | 420 | 959 | 1095 | 1142 | 1936 | 1940 | 3125 | 3318 | 3381 |
| 8 | 814 | 821 | 1531 | 1648 | 1750 | 2799 | 2816 | 4088 | 4347 | 4403 |
| 7 | 1317 | 1334 | 2359 | 2591 | 2656 | 4019 | 4026 | 5205 | 5327 | 5406 |
| 6 | 2112 | 2121 | 3543 | 3853 | 3957 | 5648 | 5640 | 6551 | 6815 | 6902 |
| 5 | 3722 | 3742 | 5494 | 5808 | 5970 | 7634 | 7639 | 8369 | 8417 | 8477 |
| 4 | 6383 | 6387 | 8164 | 8522 | 8635 | 9943 | 9946 | 9677 | 9589 | 9505 |
| 3 | 10177 | 10190 | 11699 | 11862 | 11893 | 12315 | 12317 | 10981 | 10656 | 10472 |
| 2 | 16101 | 16123 | 15693 | 15501 | 15402 | 13832 | 13831 | 11166 | 10632 | 10454 |
| 1 | 21626 | 21598 | 18446 | 17665 | 17389 | 13240 | 13212 | 10142 | 9519 | 9259 |
| 0 | 23182 | 23115 | 16900 | 15947 | 15574 | 10783 | 10751 | 8151 | 7684 | 7520 |

Table C.1: Pair statistics of Wiki-dictionaries without comments in brakets on the UMC train set

| | NUMBER OF SENTENCES | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (COUNTS BY WORD FORMS) | | | | | (COUNTS BY NORMAL FORMS) | | | | |
| # of pairs | I (EN-DE-RU) | I+R (EN-DE-RU) | I | I+R | I +R +T | I (EN-DE-RU) | I+R (EN-DE-RU) | I | I+R | I +R +T |
| 23 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 21 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 2 |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 1 |
| 16 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 |
| 15 | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 2 | 4 | 3 |
| 14 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 2 | 3 | 3 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 1 | 2 |
| 12 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 3 | 4 | 6 |
| 11 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 10 | 10 | 11 |
| 10 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 9 | 15 | 12 |
| 9 | 2 | 2 | 3 | 5 | 5 | 5 | 5 | 15 | 12 | 12 |
| 8 | 2 | 2 | 4 | 5 | 5 | 17 | 18 | 27 | 32 | 35 |
| 7 | 3 | 4 | 13 | 12 | 12 | 28 | 27 | 47 | 44 | 48 |
| 6 | 12 | 11 | 18 | 20 | 24 | 37 | 37 | 47 | 60 | 59 |
| 5 | 15 | 15 | 36 | 40 | 38 | 64 | 64 | 71 | 75 | 76 |
| 4 | 48 | 49 | 66 | 74 | 73 | 81 | 83 | 100 | 89 | 86 |
| 3 | 91 | 90 | 104 | 105 | 109 | 128 | 127 | 104 | 110 | 107 |
| 2 | 140 | 142 | 161 | 166 | 163 | 137 | 139 | 116 | 112 | 111 |
| 1 | 219 | 221 | 187 | 176 | 174 | 130 | 129 | 115 | 109 | 108 |
| 0 | 216 | 212 | 155 | 144 | 144 | 112 | 110 | 74 | 65 | 65 |

Table C.2: Pair statistics of Wiki-dictionaries without comments in brakets on the UMC development set

| | NUMBER OF SENTENCES | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (COUNTS BY WORD FORMS) | | | | | (COUNTS BY NORMAL FORMS) | | | | |
| # of pairs | I (EN-DE-RU) | I+R (EN-DE-RU) | I | I+R | I +R +T | I (EN-DE-RU) | I+R (EN-DE-RU) | I | I+R | I +R +T |
| 22 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 3 |
| 15 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 7 | 5 | 6 |
| 14 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 3 | 6 | 6 |
| 13 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 11 | 11 | 12 |
| 12 | 1 | 1 | 2 | 4 | 3 | 7 | 7 | 10 | 7 | 8 |
| 11 | 0 | 0 | 2 | 0 | 1 | 6 | 7 | 13 | 16 | 16 |
| 10 | 1 | 1 | 6 | 8 | 7 | 9 | 8 | 20 | 26 | 26 |
| 9 | 4 | 4 | 9 | 9 | 12 | 13 | 14 | 26 | 29 | 31 |
| 8 | 7 | 7 | 12 | 15 | 13 | 33 | 32 | 36 | 43 | 43 |
| 7 | 9 | 9 | 17 | 19 | 19 | 34 | 34 | 58 | 58 | 63 |
| 6 | 16 | 16 | 32 | 32 | 33 | 65 | 65 | 88 | 94 | 100 |
| 5 | 37 | 38 | 40 | 52 | 56 | 82 | 87 | 94 | 92 | 86 |
| 4 | 58 | 58 | 90 | 95 | 100 | 141 | 136 | 132 | 133 | 136 |
| 3 | 129 | 131 | 147 | 149 | 148 | 154 | 156 | 127 | 123 | 119 |
| 2 | 183 | 181 | 200 | 196 | 200 | 158 | 156 | 138 | 129 | 126 |
| 1 | 275 | 276 | 250 | 242 | 236 | 161 | 161 | 144 | 139 | 136 |
| 0 | 280 | 278 | 193 | 179 | 171 | 133 | 133 | 91 | 84 | 81 |

Table C.3: Pair statistics of Wiki-dictionaries without comments in brakets on the UMC test set